PhD Thesis Defense

ROBUST EFFICIENT EDGE AI: New Principles and Frameworks for Empowering AI on Edge Devices

Rahul Duggal
Computer Science PhD Student
Georgia Institute of Technology

July 14, 2022

Thesis Committee

Dr. Polo Chau (Advisor)
School of CSE
Georgia Institute of Technology

Dr. Richard Vuduc
School of CSE
Georgia Institute of Technology

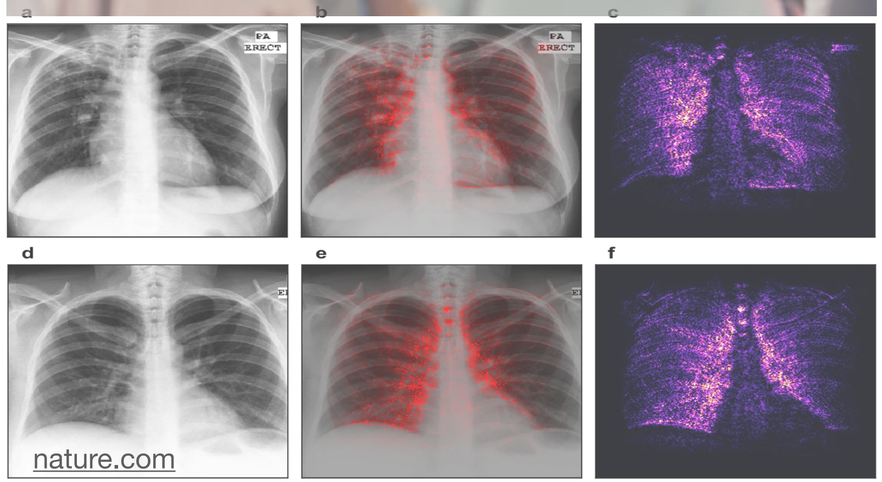
Dr. Callie Hao
School of ECE
Georgia Institute of Technology

Dr. Jimeng Sun (Co-advisor)
Dept. of CS
University of Illinois at Urbana-Champaign

Dr. Ada Gavrilovska
School of CS
Georgia Institute of Technology

Impact of AI Today

towardsdatascience.com

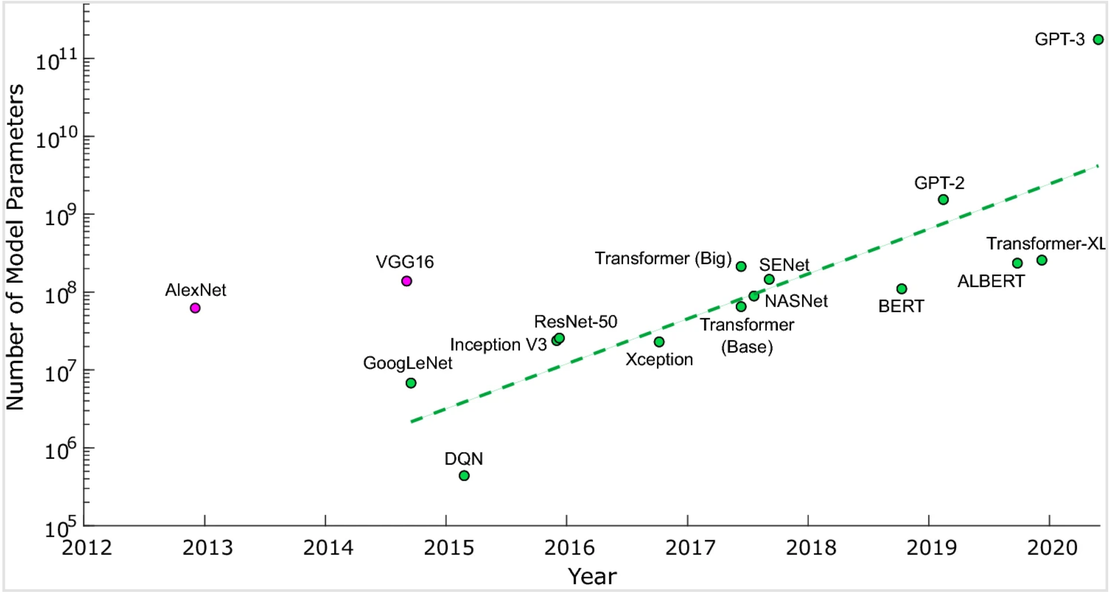


nature.com

re-work blog

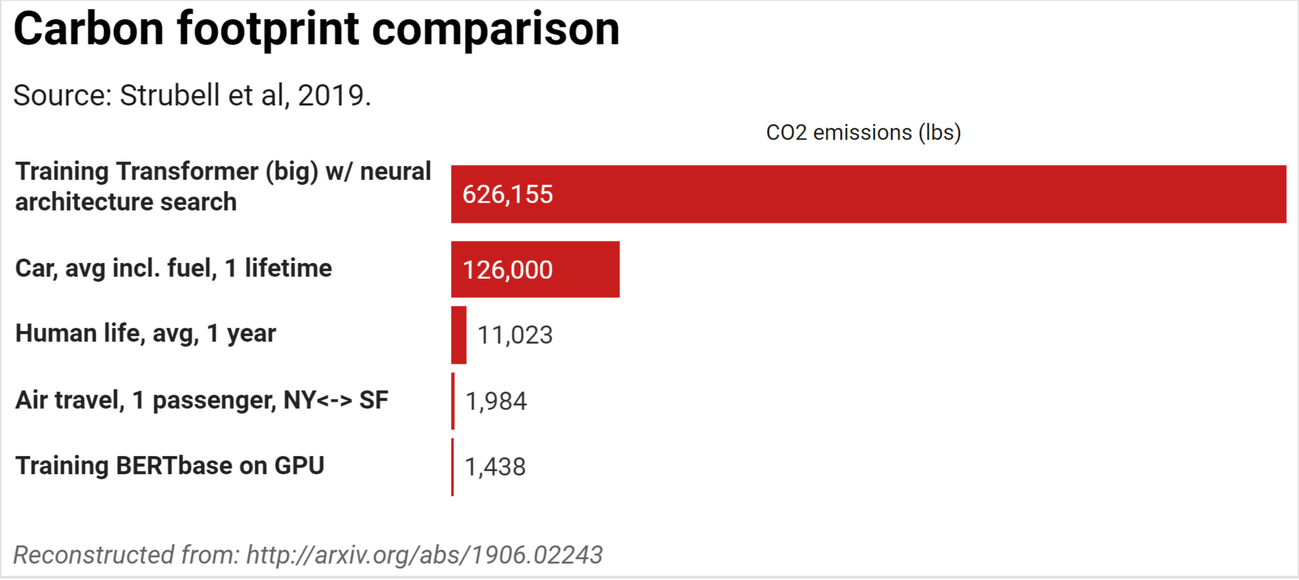
AI models are growing larger

More parameters



Bernstein, L., Sludds, A., Hamerly, R. *et al.* Freely scalable and reconfigurable optical hardware for deep learning. *Sci Rep* **11,** 3144 (2021). https://doi.org/10.1038/s41598-021-82543-3

Increasing energy consumption

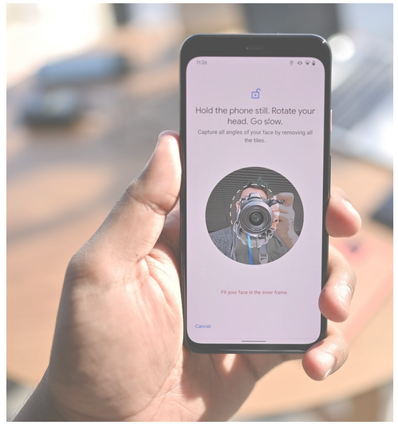


Edge AI Now Important

Huge Data-centers

Edge Devices

Health monitoring



Face Unlock

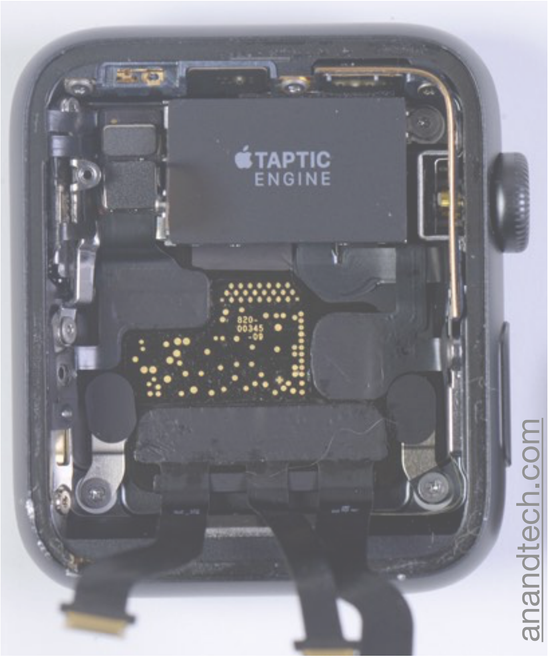
- ✓ Privacy
- ✓ Low Latency
- ✓ No Network

How to extend deployment frontier of AI to **edge devices**?

Edge AI Research Challenges

1. Efficiency

Small GPU/CPU



Limits the compute footprint of DNNs

Small battery

Limits energy consumption of DNNs

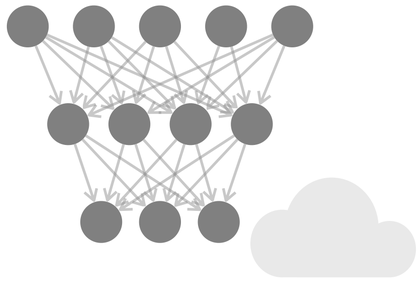
2. **Robustness** against Noise

Noisy data from **low-power sensors**

Noisy **natural environments**

PhD Research Overview

Robust Efficient Edge AI



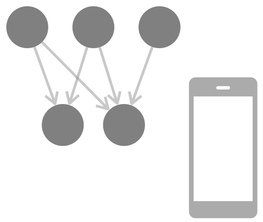
Large models deployed on cloud servers

Efficient edge AI

- ★ **CMP-NAS** CVPR 21
- ★ **CUP** Big Data 21
- ★ **REG-NAS** ECCV 22

Robust edge AI

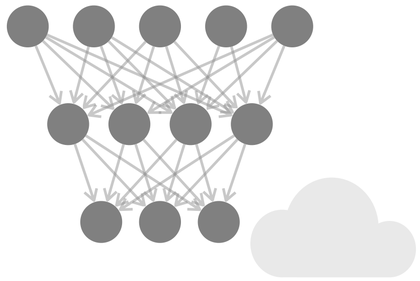
- ★ **REST** WWW 20
- ★ **HAR** Big Data 21
- ★ **IMB-NAS** Submitted, AAAI 23
- ★ **MalNet** Submitted, CIKM 22
- ★ **Neurocarto** SIGGRAPH 22
- ★ **CptEvo** Arxiv



Small models deployed on edge devices

Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Efficient edge AI

★ **CMP-NAS**
CVPR 21

★ **CUP**
Big Data 21

★ **REG-NAS**
ECCV 22

Robust edge AI

★ **REST**
WWW 20

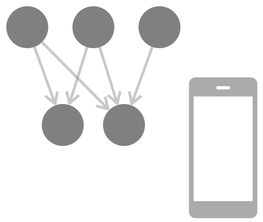
★ **HAR**
Big Data 21

★ **IMB-NAS**
Submitted, AAAI 23

★ **MalNet**
Submitted, CIKM 22

★ **Neurocarto**
SIGGRAPH 22

★ **CptEvo**
Arxiv



Small models deployed on edge devices

Thesis Overview

Robust Efficient Edge AI

Part I: New frameworks for **efficient edge AI**

CMP-NAS Bottom-up design using neural architecture search
[CVPR 21]

CUP Top-down design using pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

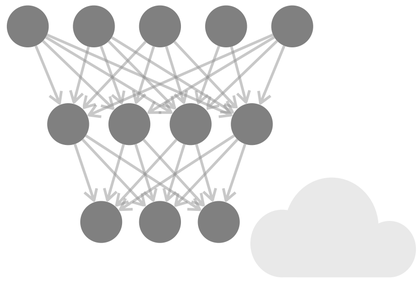
REST Test time robustness with pruning
[WWW 20]

HAR Training time robustness with early exiting
[Big Data 21]

IMB-NAS Neural architecture search under class imbalance
[Submitted to AAAI 2023]

Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

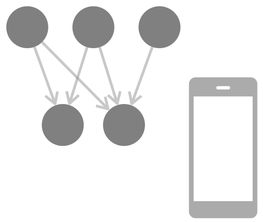
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

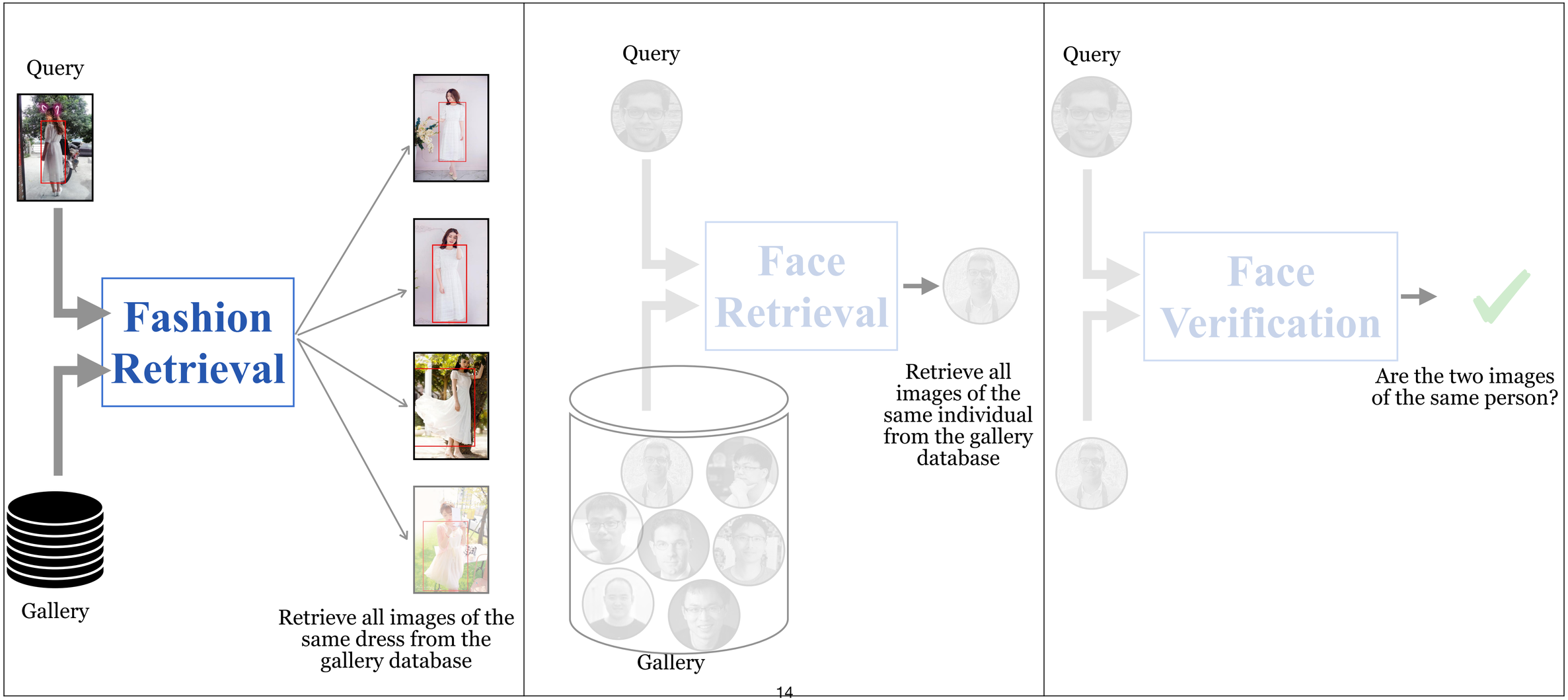
Collaboration with amazon | science

Compatibility-Aware Heterogeneous Visual Search

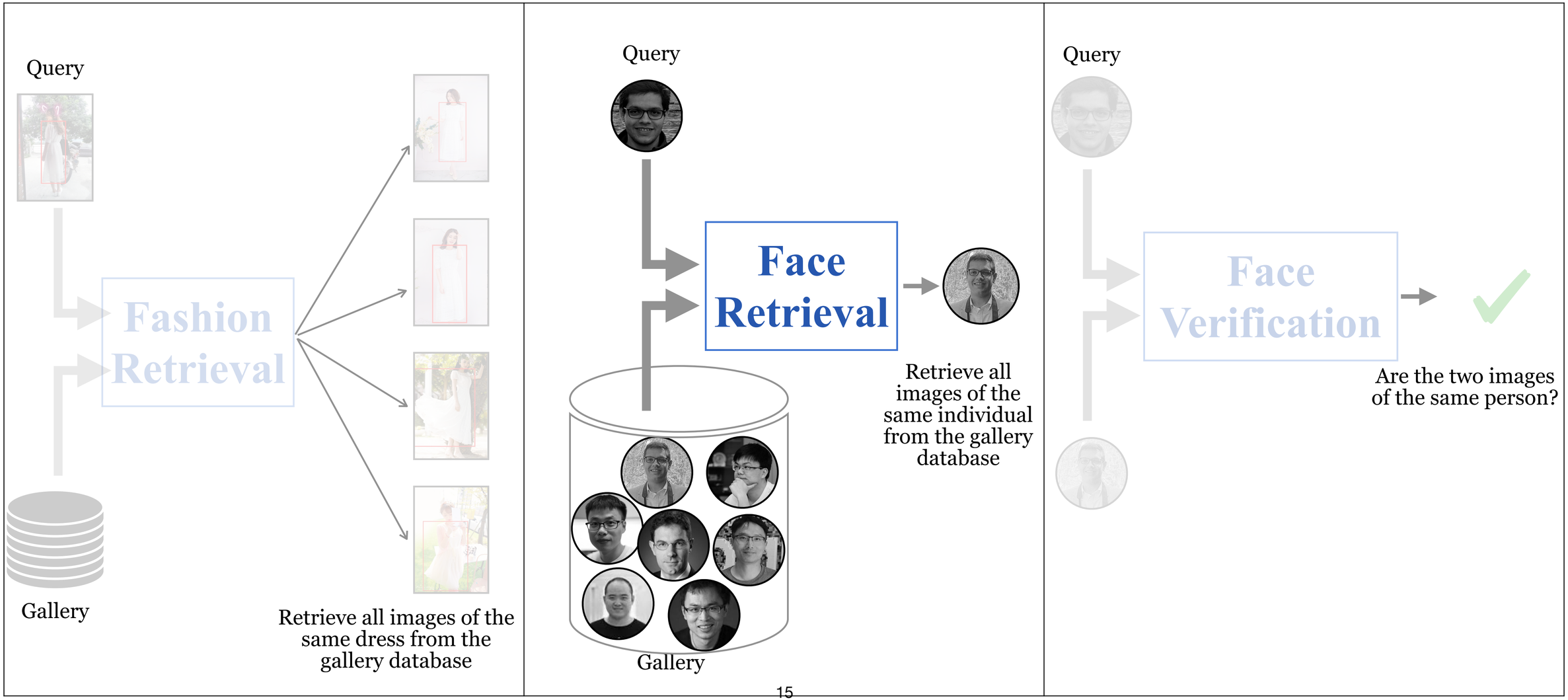
Rahul Duggal, Hao Zou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, Stefano Soatto

IEEE Computer Vision and Patten Recognition **(CVPR)**, 2021

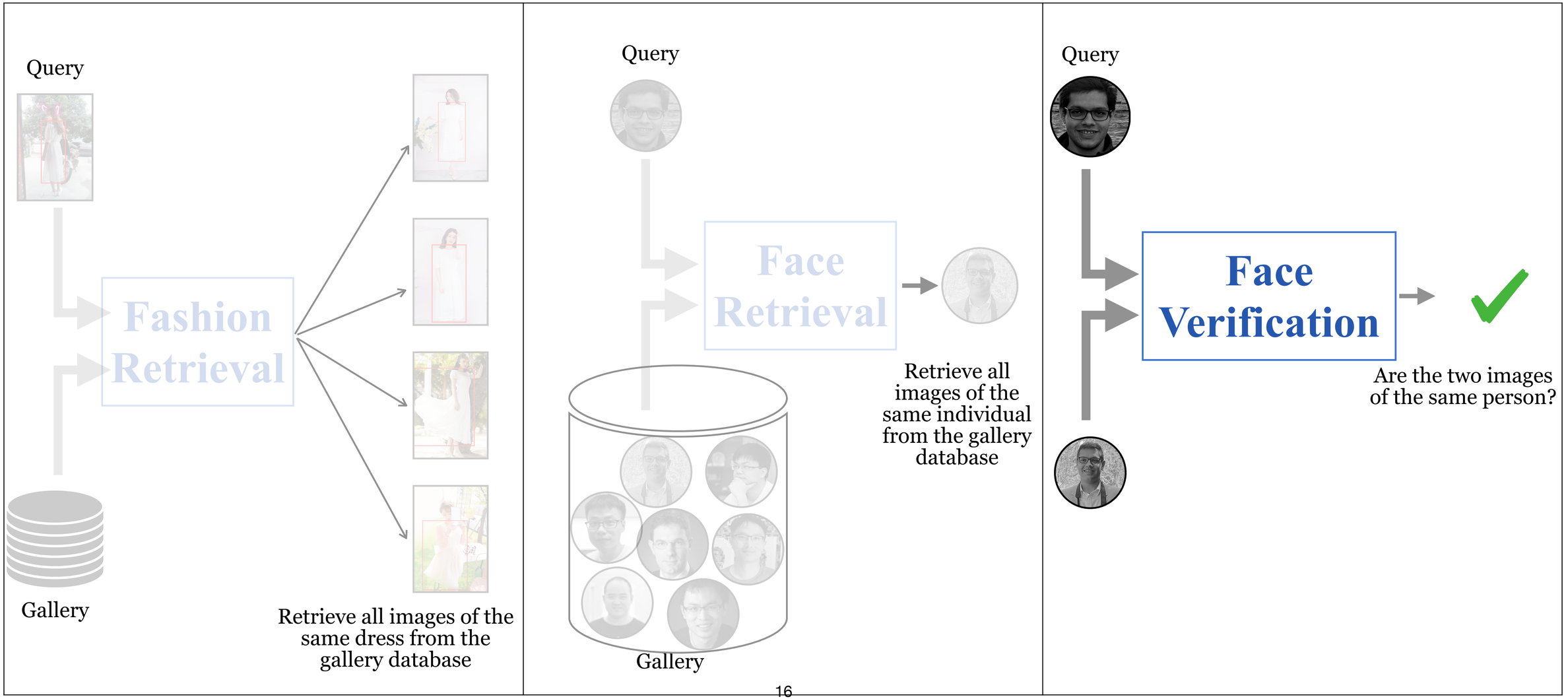
Visual Search



Visual Search



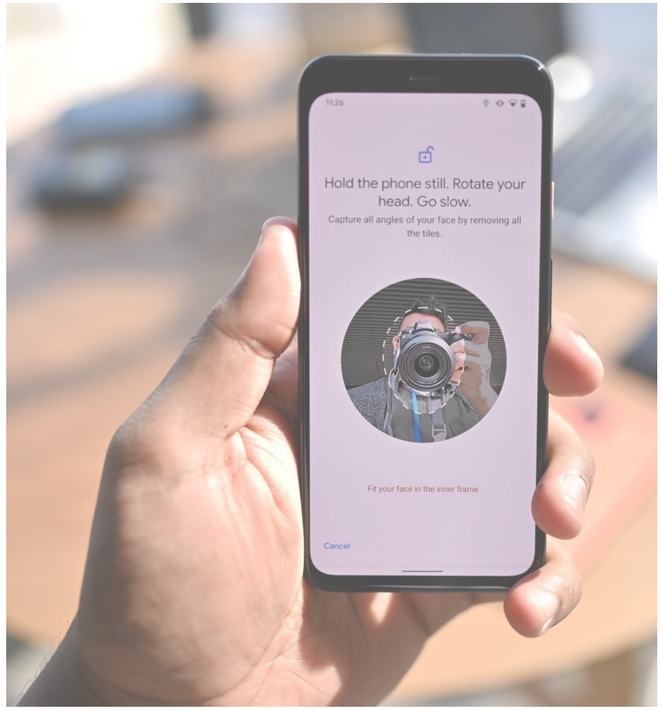
Visual Search



Our Goal

Visual search on the edge

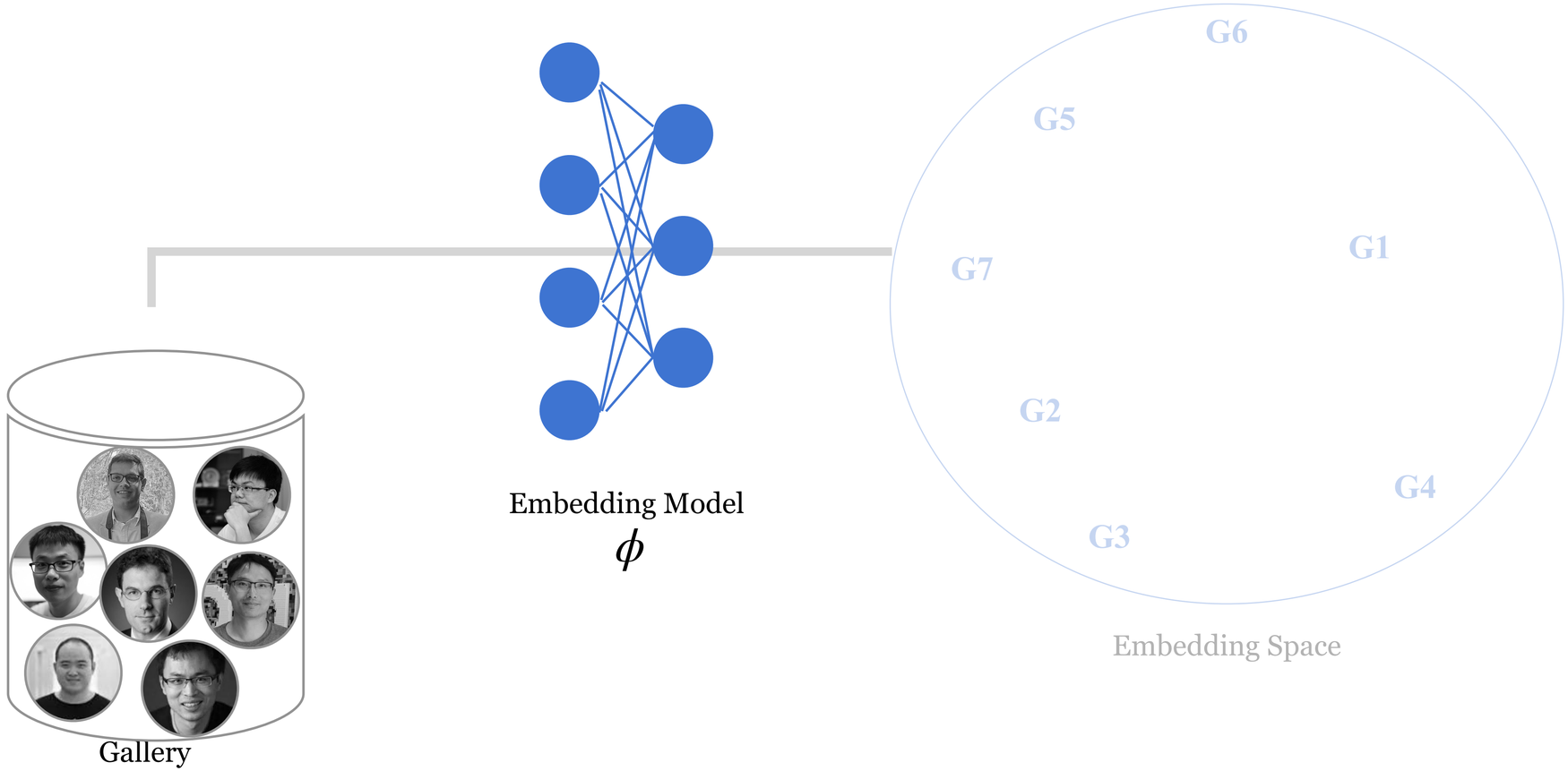
Face Unlock on Mobile Phone



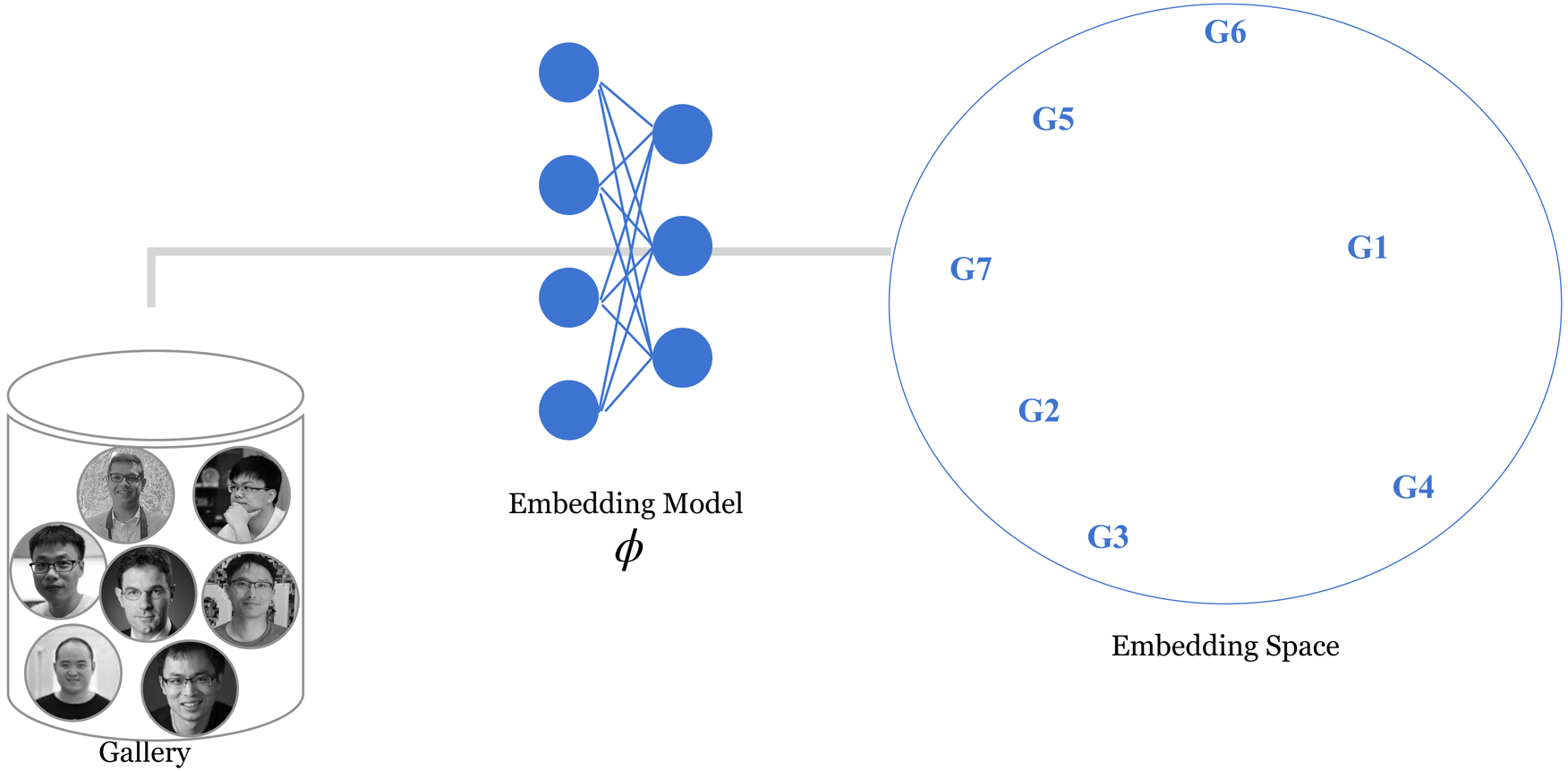
Fashion retrieval on a mobile phone

Hadi Kiapour, M., et al. "Where to buy it: Matching street clothing photos in online shops." *Proceedings of the IEEE international conference on computer vision*. 2015.

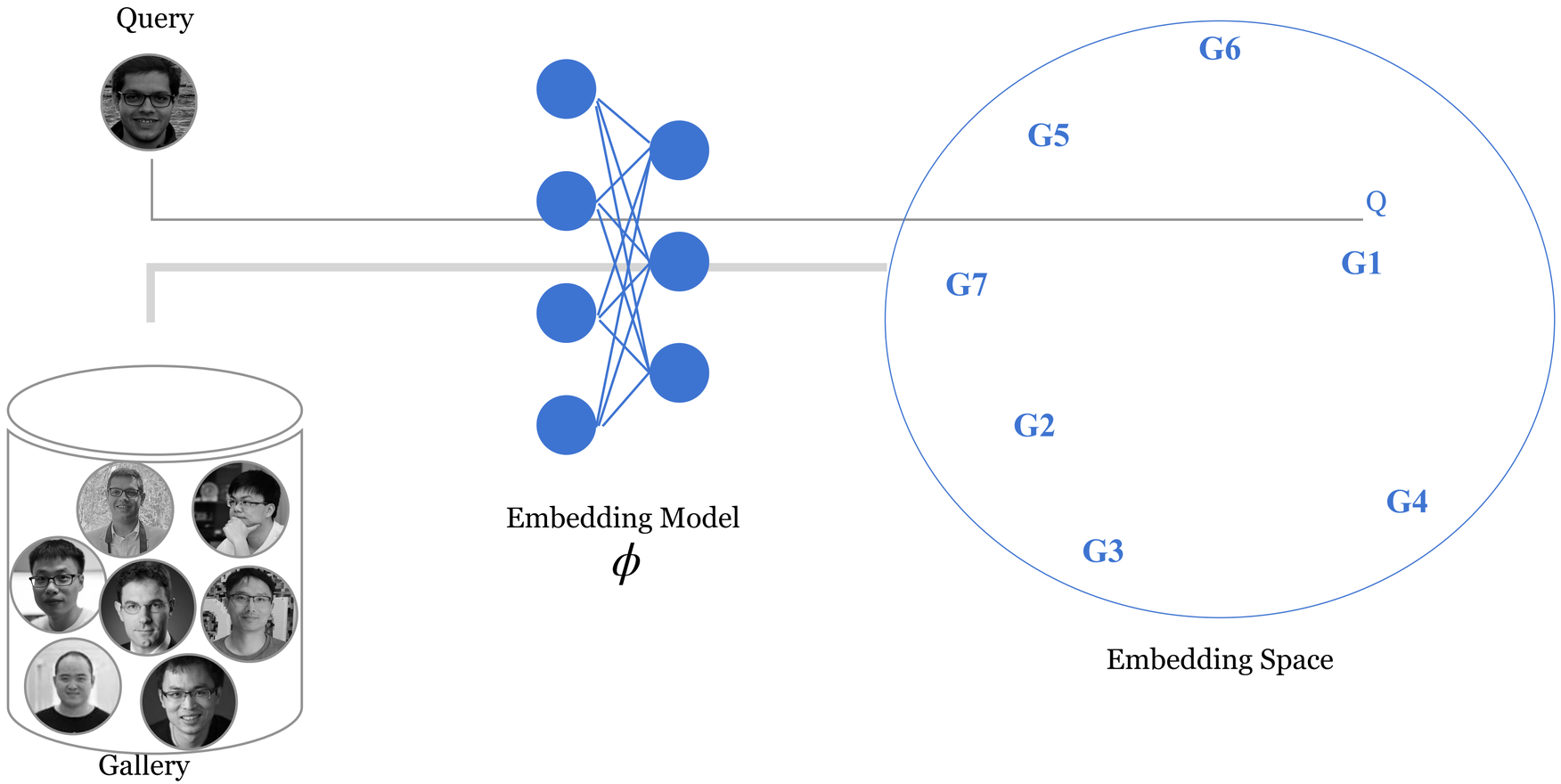
Visual Search Pipeline



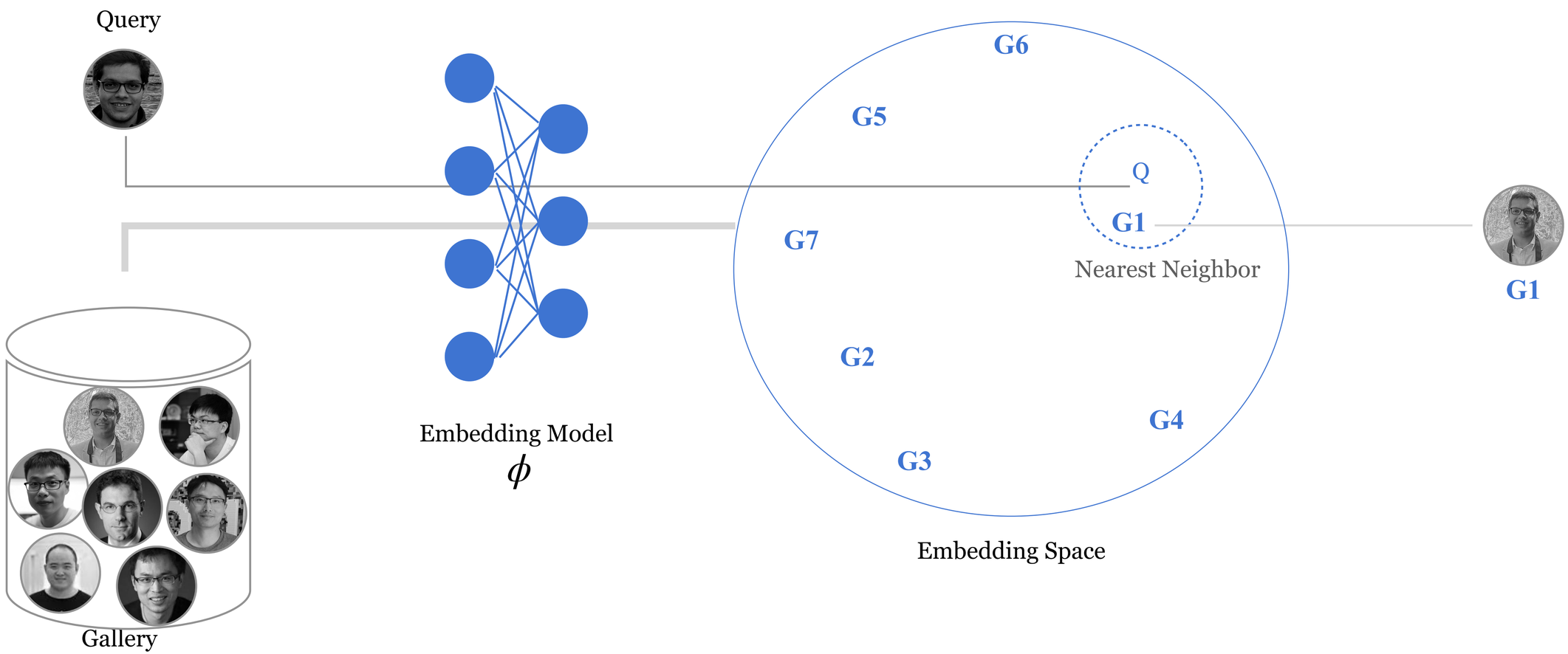
Visual Search Pipeline



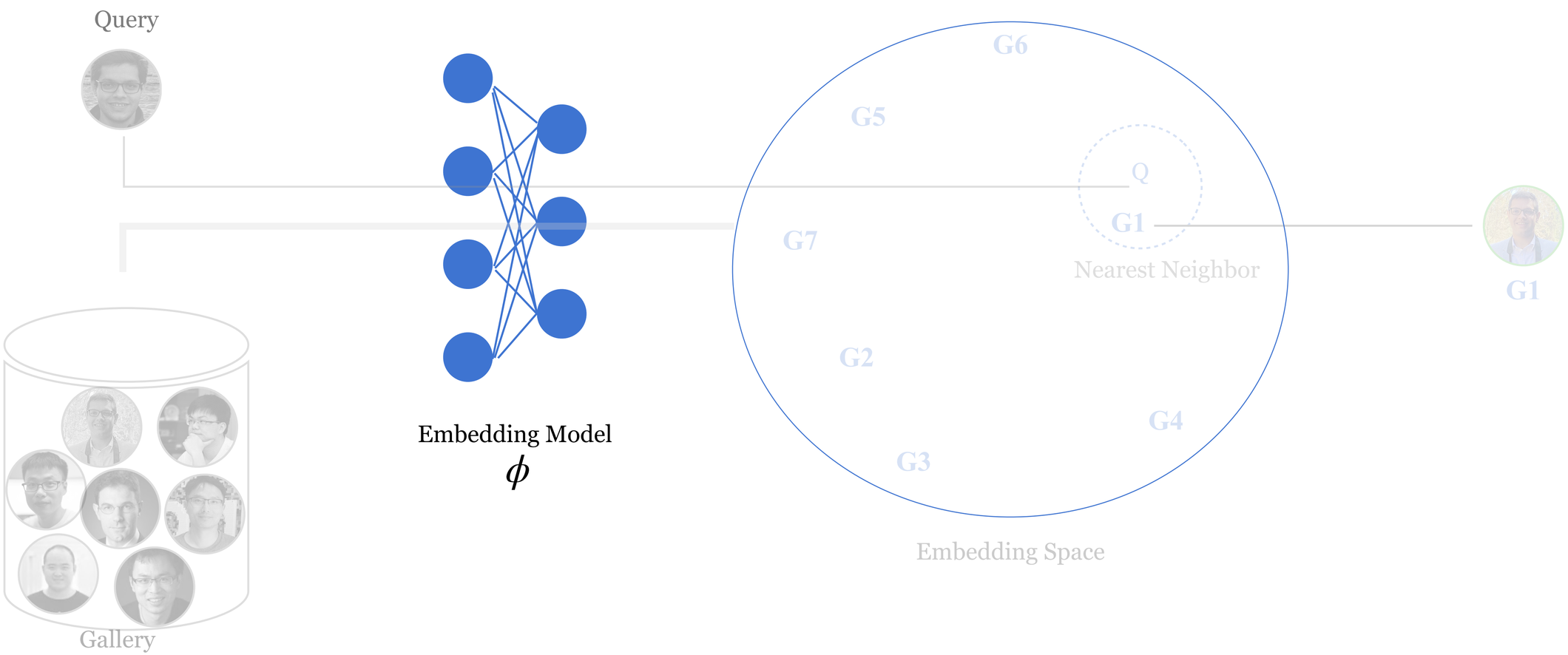
Visual Search Pipeline



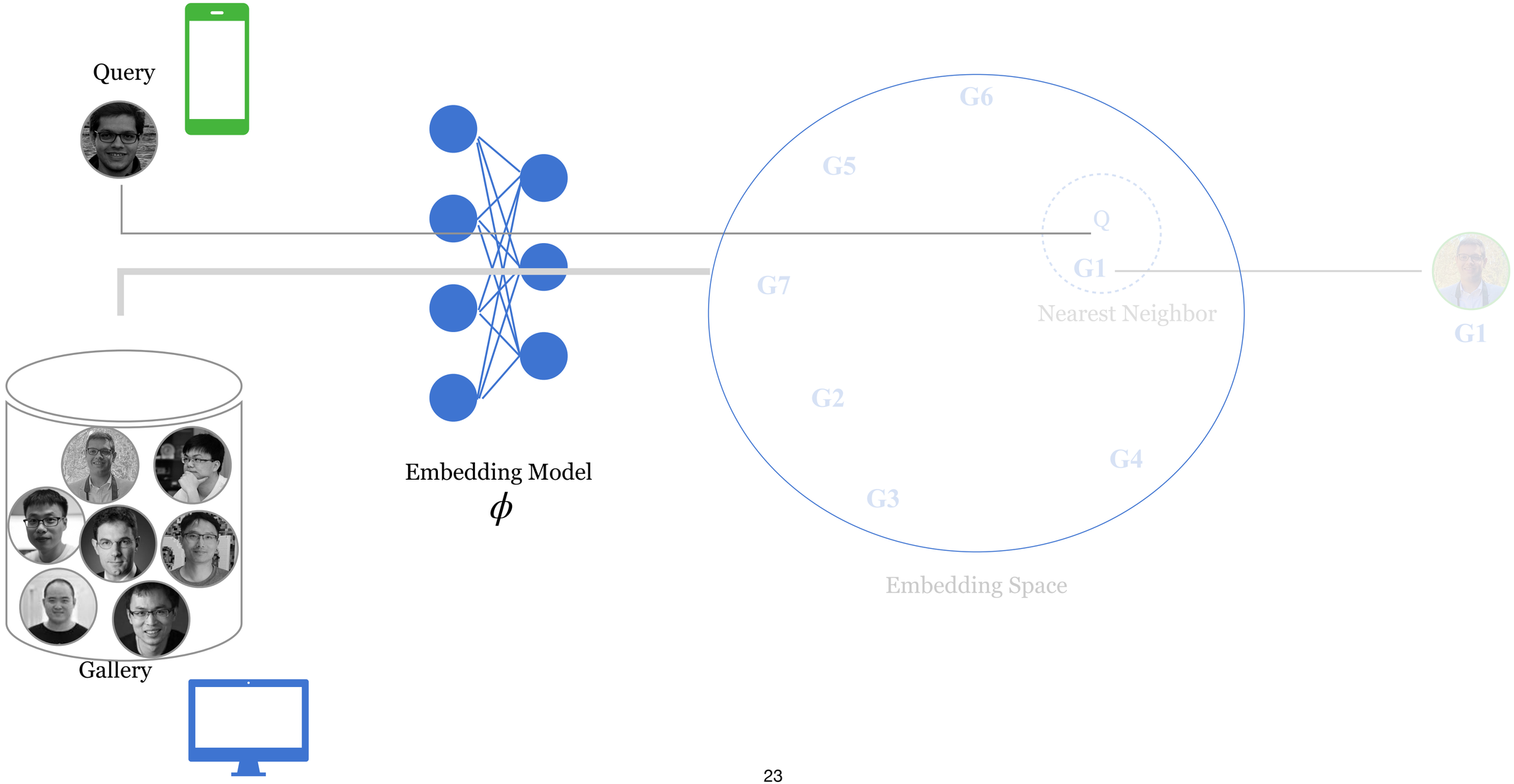
Visual Search Pipeline



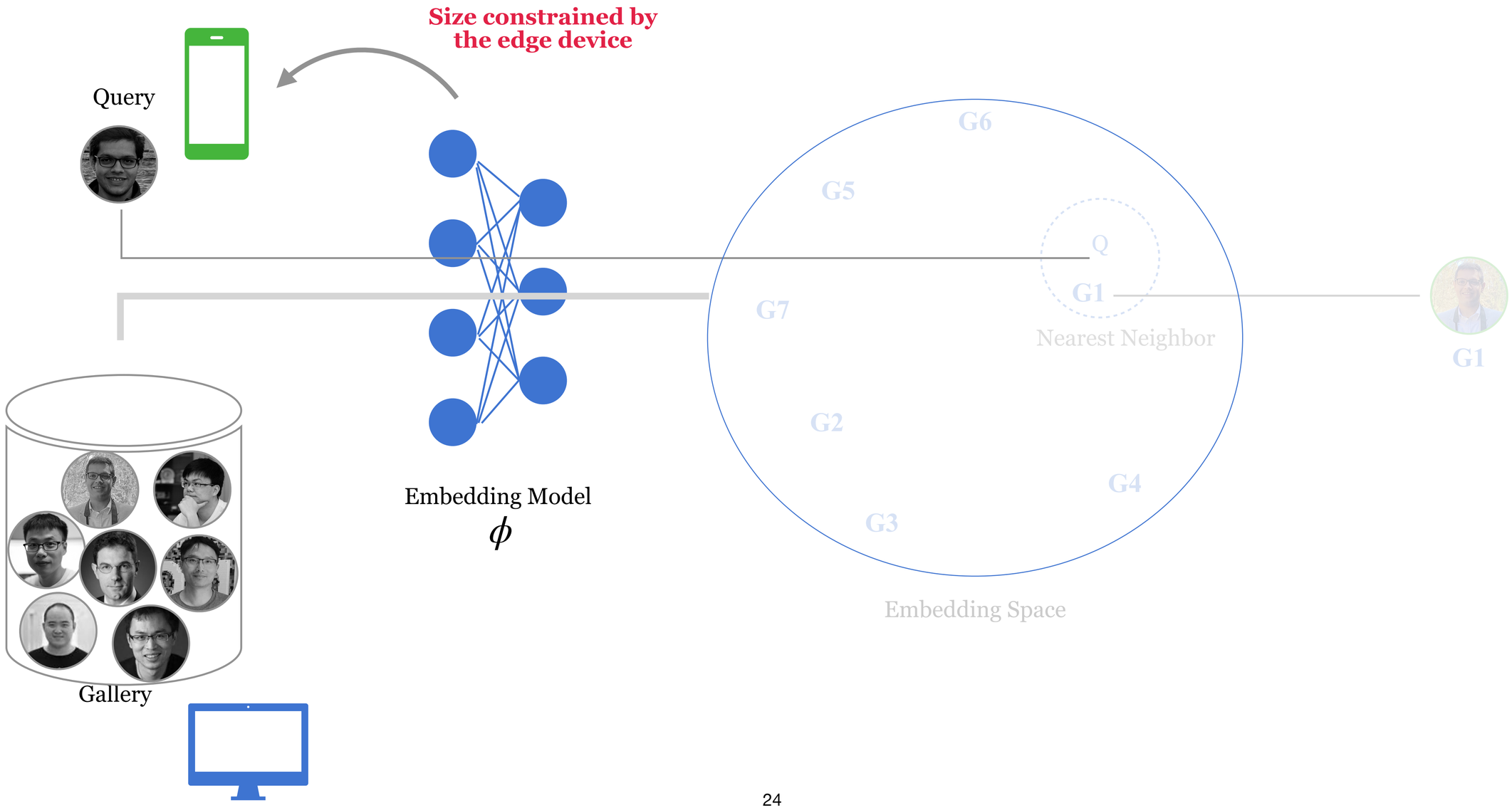
Visual Search Pipeline



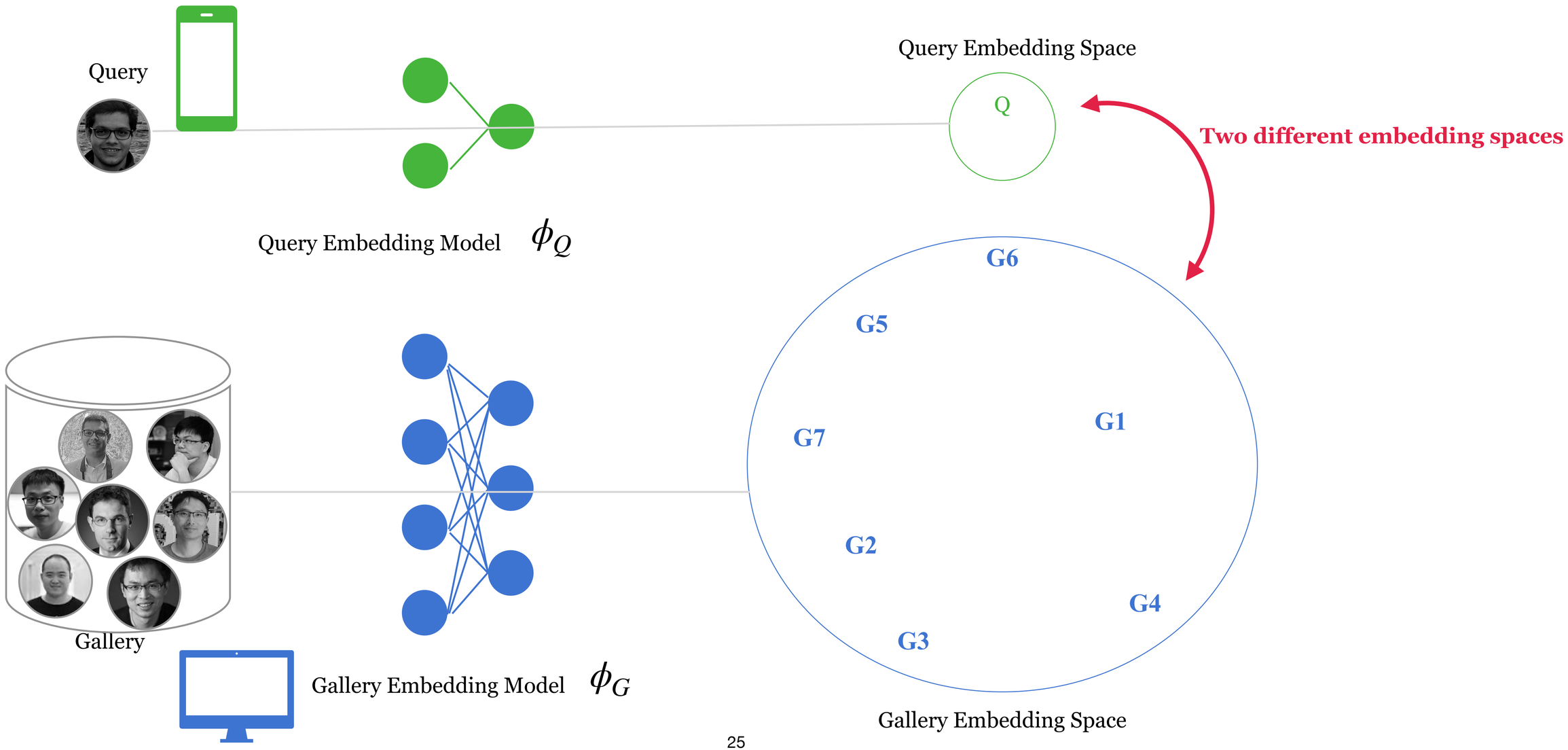
Visual Search Pipeline



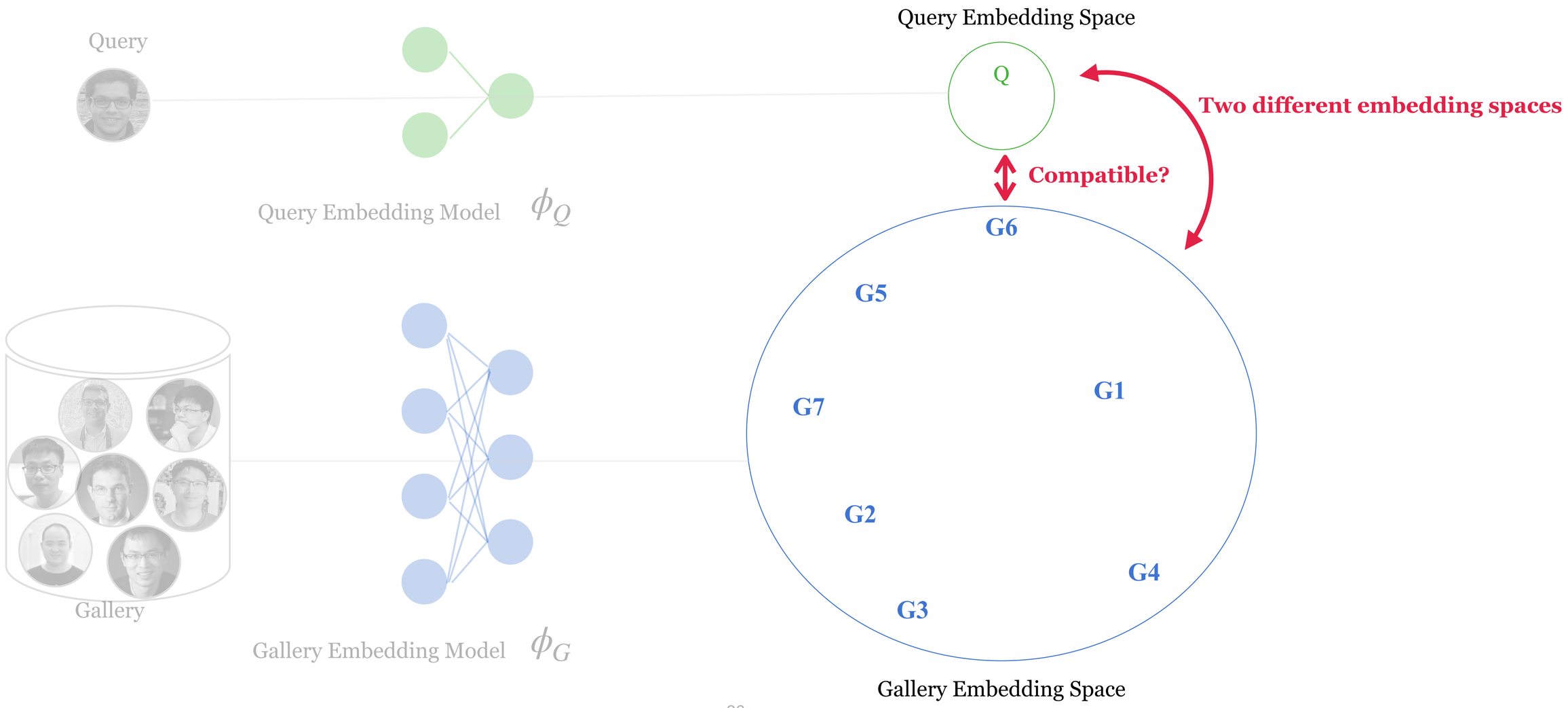
Visual Search Pipeline



Heterogeneous Visual Search



Heterogeneous Visual Search



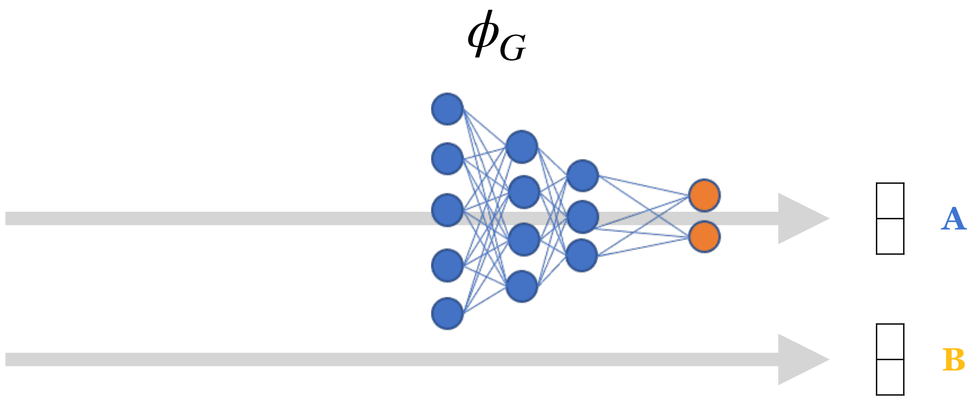
Two forms of compatibility

1. By optimizing the query model's weights
   - Backward compatible training (BCT) [1]

2. By optimizing the query model's architecture
   - CMP-NAS (This work)

[1] Shen, Yantao, et al. "Towards backward-compatible representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

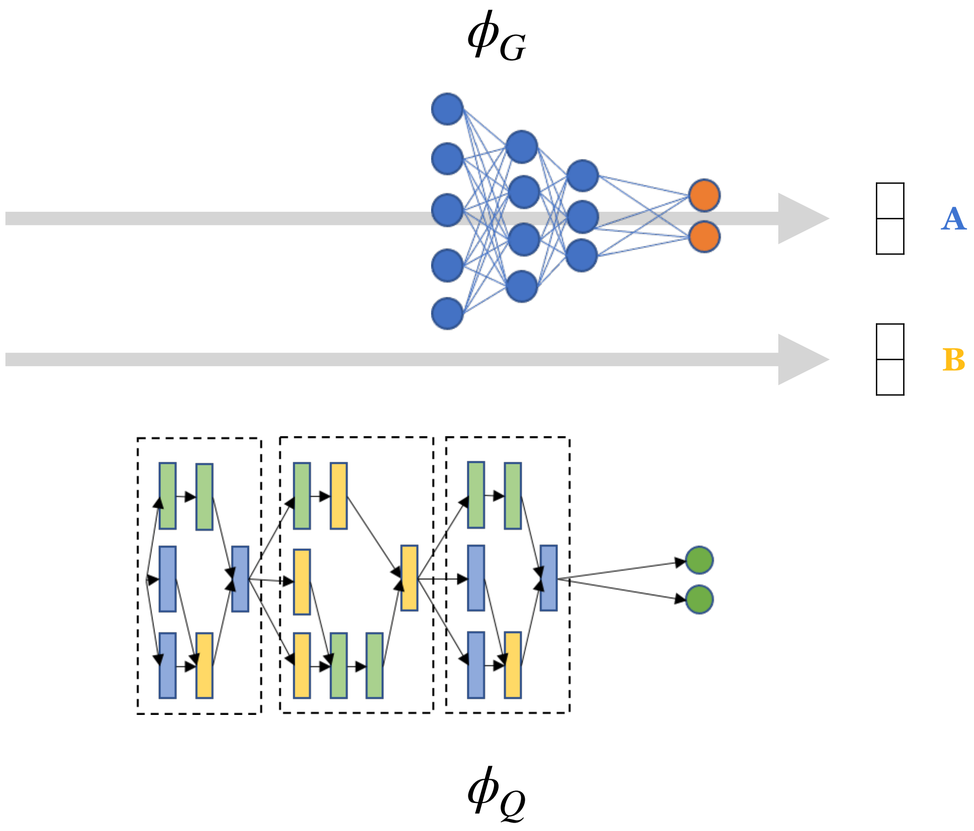
Our method
CMP-NAS



Step 1: Train ϕ_G

- $\mathscr{L}$ = CrossEntropy(**A** , **B**)

Our method
CMP-NAS

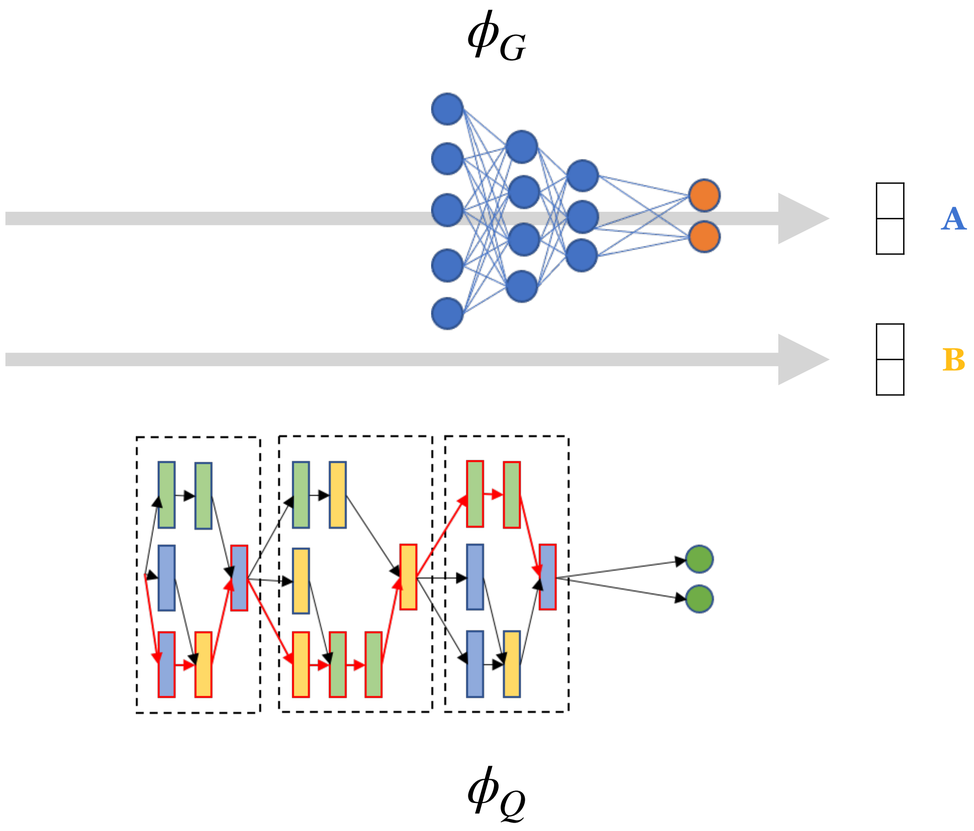


Step 1: **Train ϕ_G**

- $\mathscr{L}$ = CrossEntropy(**A** , **B**)

Step 2: **Train ϕ_Q**

Our method
CMP-NAS



Step 1: **Train ϕ_G**

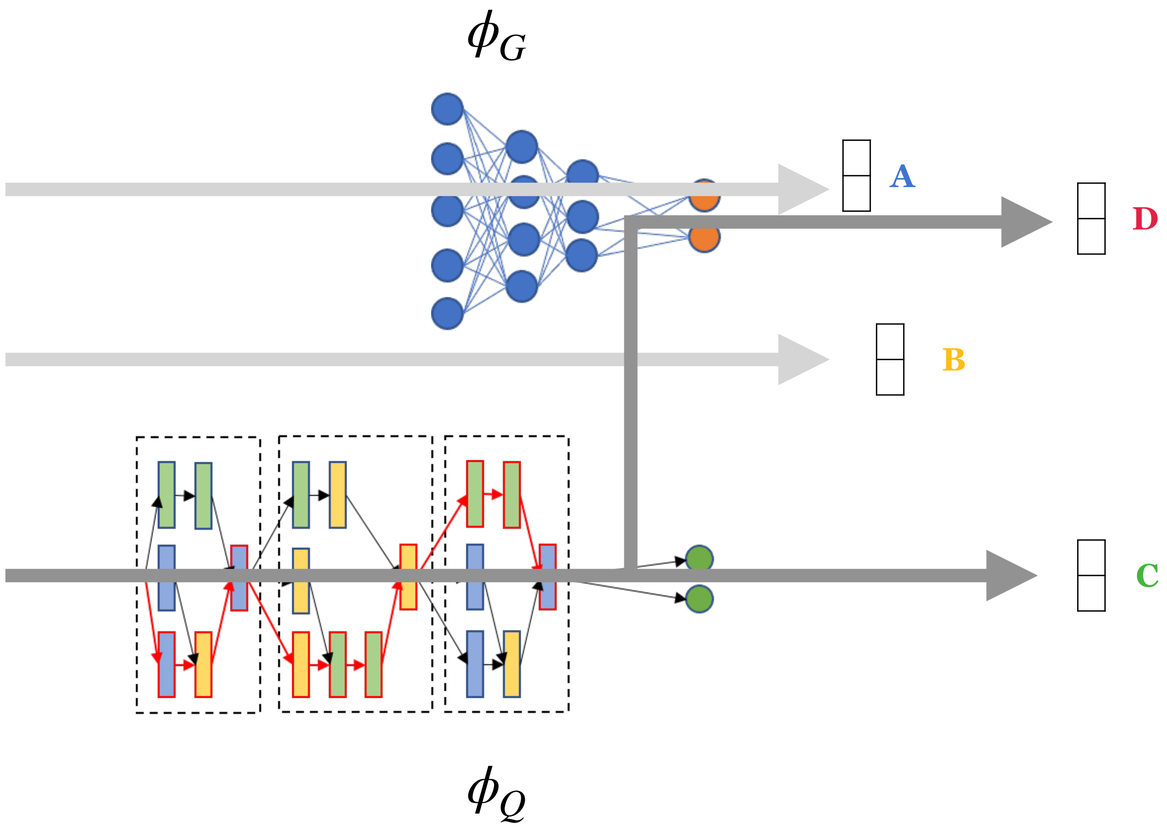
- $\mathscr{L}$ = CrossEntropy(**A** , **B**)

Step 2: **Train ϕ_Q**

1. **Sample an architecture a [2].**
   - **a** $\sim U(\Omega)$

[1] Shen, Yantao, et al. "Towards backward-compatible representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[2] Guo, Zichao, et al. "Single path one-shot neural architecture search with uniform sampling." *European Conference on Computer Vision*. Springer, Cham, 2020.

Our method
CMP-NAS



Step 1: **Train ϕ_G**

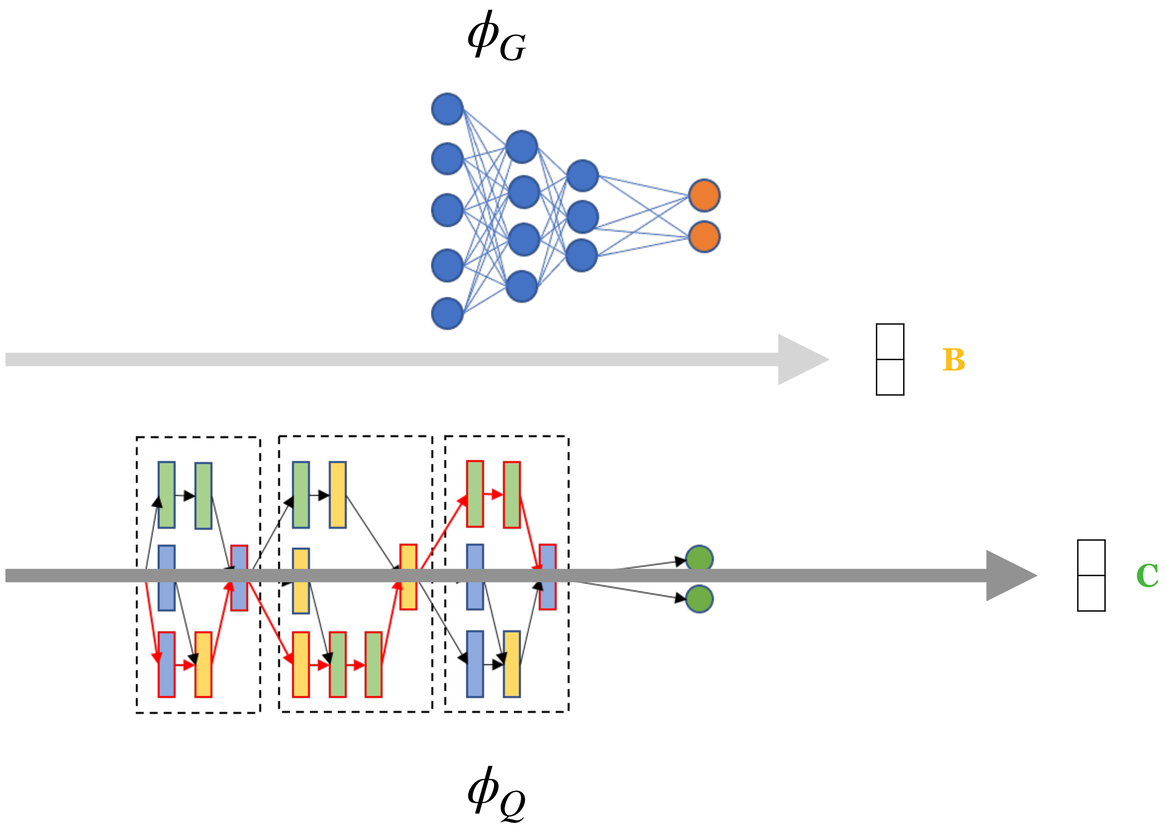
- $\mathscr{L}$ = CrossEntropy(**A** , **B**)

Step 2: **Train ϕ_Q**

1. **Sample an architecture a [2].**
   - **a** $\sim U(\Omega)$
2. **Minimize the BCT Loss [1] for a**
   - $\mathscr{L}$ = CrossEntropy(**C** , **B**) + CrossEntropy(**D** , **C**)

[1] Shen, Yantao, et al. "Towards backward-compatible representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[2] Guo, Zichao, et al. "Single path one-shot neural architecture search with uniform sampling." *European Conference on Computer Vision*. Springer, Cham, 2020.

Our method
CMP-NAS



Step 1: **Train ϕ_G**

- $\mathscr{L}$ = CrossEntropy(**A** , **B**)

Step 2: **Train ϕ_Q**

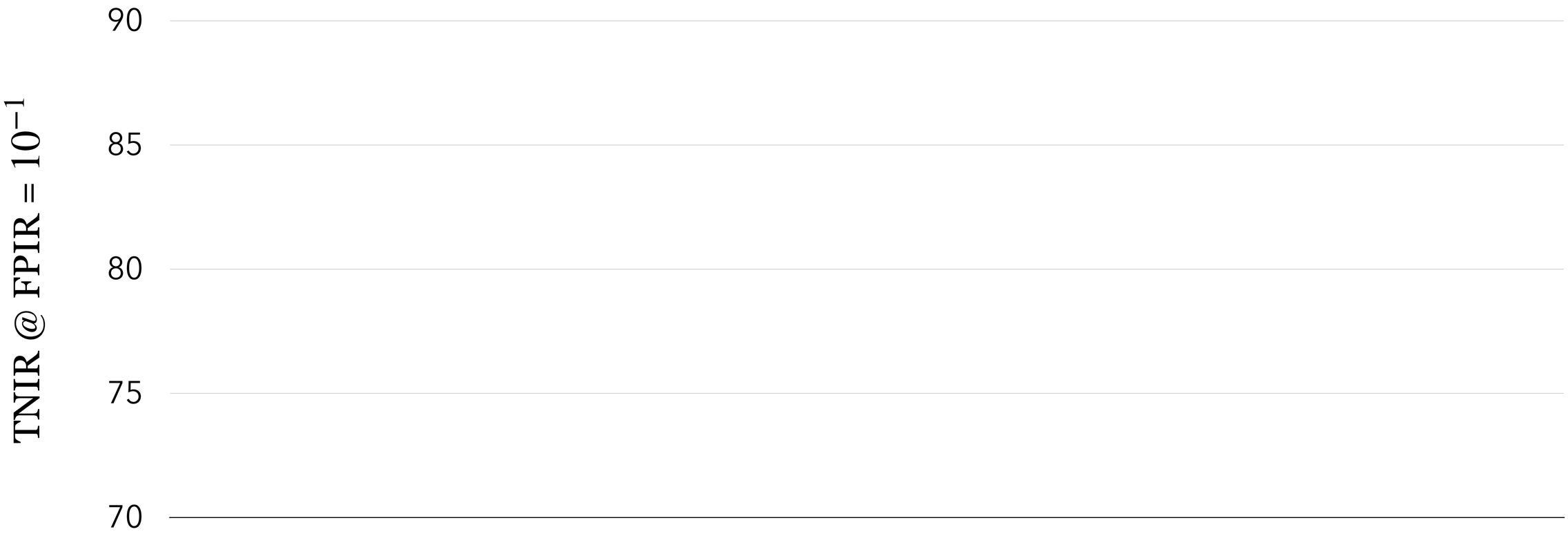
1. **Sample an architecture a [2].**
   - **a** $\sim U(\Omega)$
2. **Minimize the BCT Loss [1] for a**
   - $\mathscr{L}$ = CrossEntropy(**C** , **B**) + CrossEntropy(**D** , **B**)

Step 3: **Search for the optimal query architecture a^***

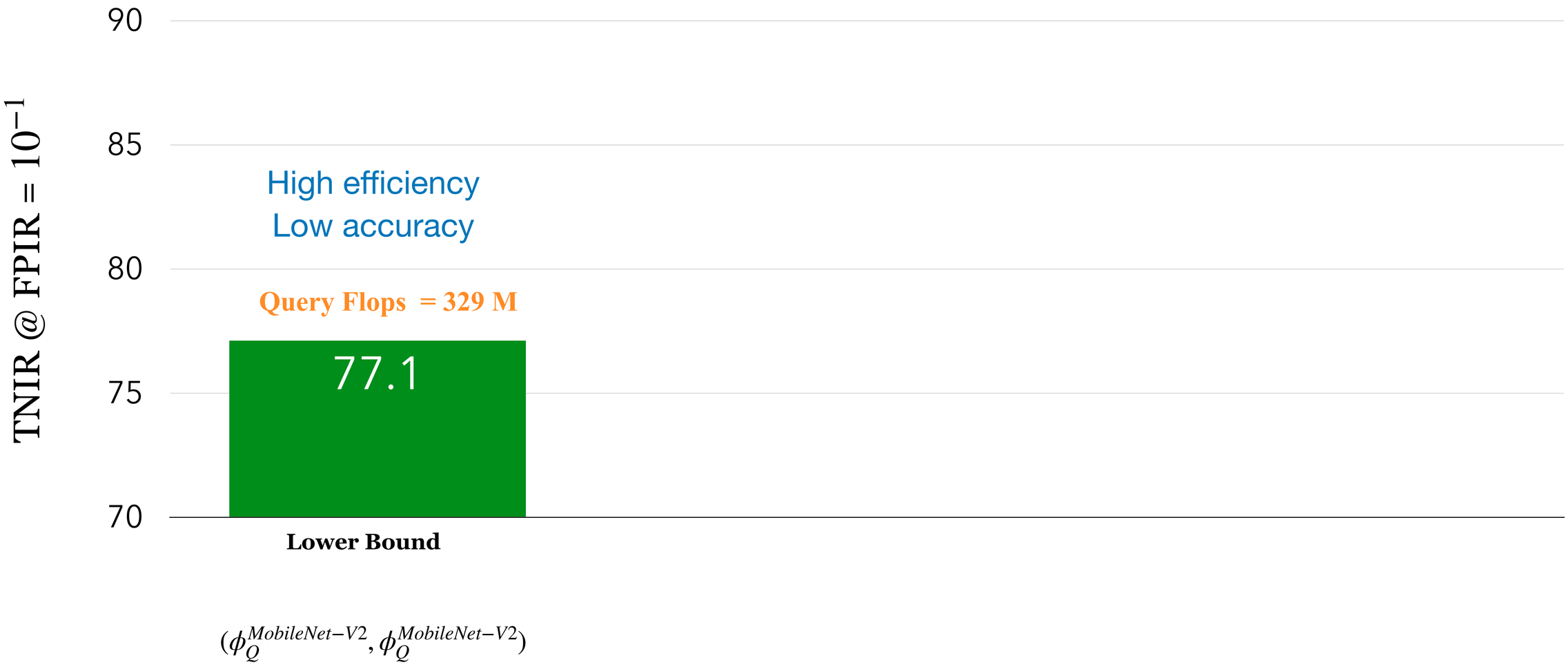
1. **Maximize reward**
   - $a^* = argmax_{a \in \Omega}(\mathscr{R})$

[1] Shen, Yantao, et al. "Towards backward-compatible representation learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[2] Guo, Zichao, et al. "Single path one-shot neural architecture search with uniform sampling." *European Conference on Computer Vision*. Springer, Cham, 2020.

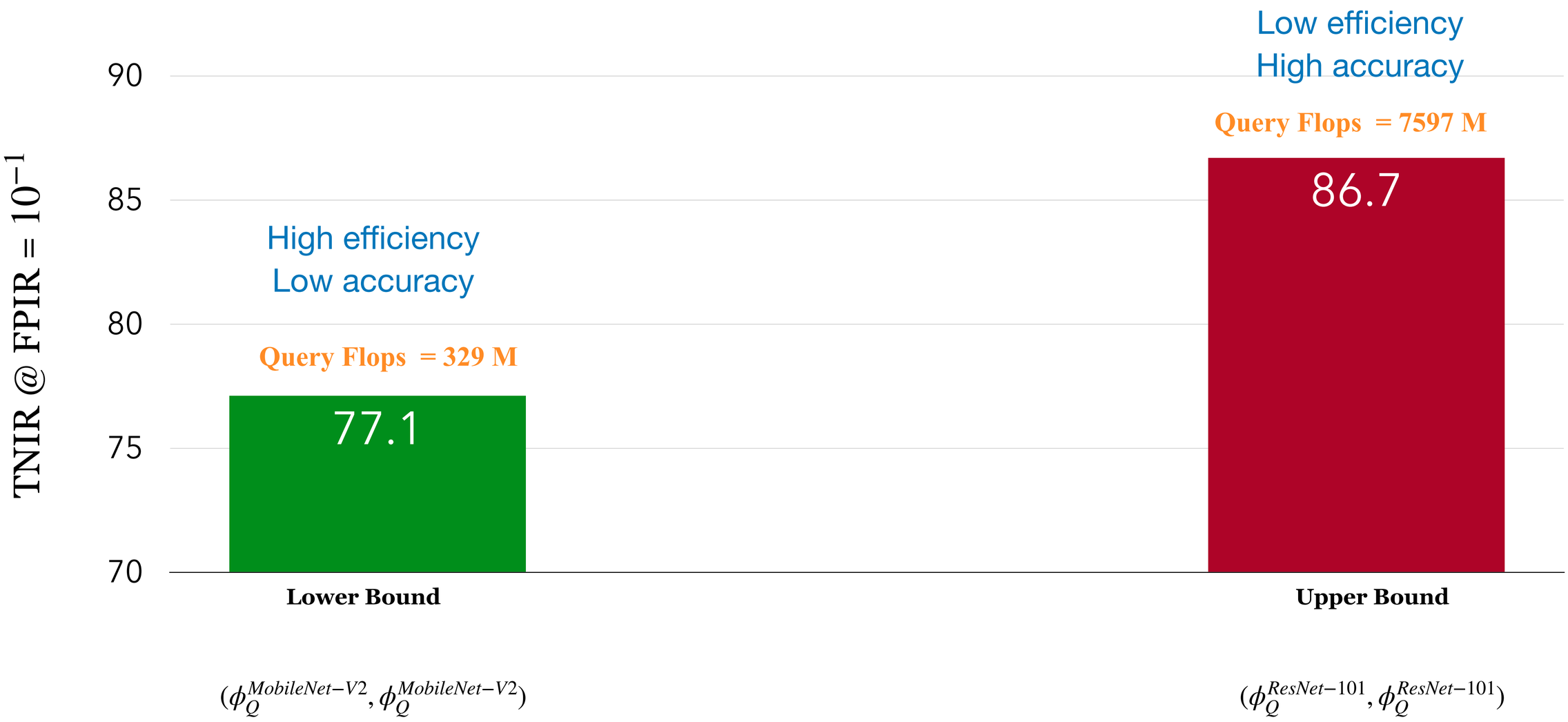
Results-1 (Face retrieval)



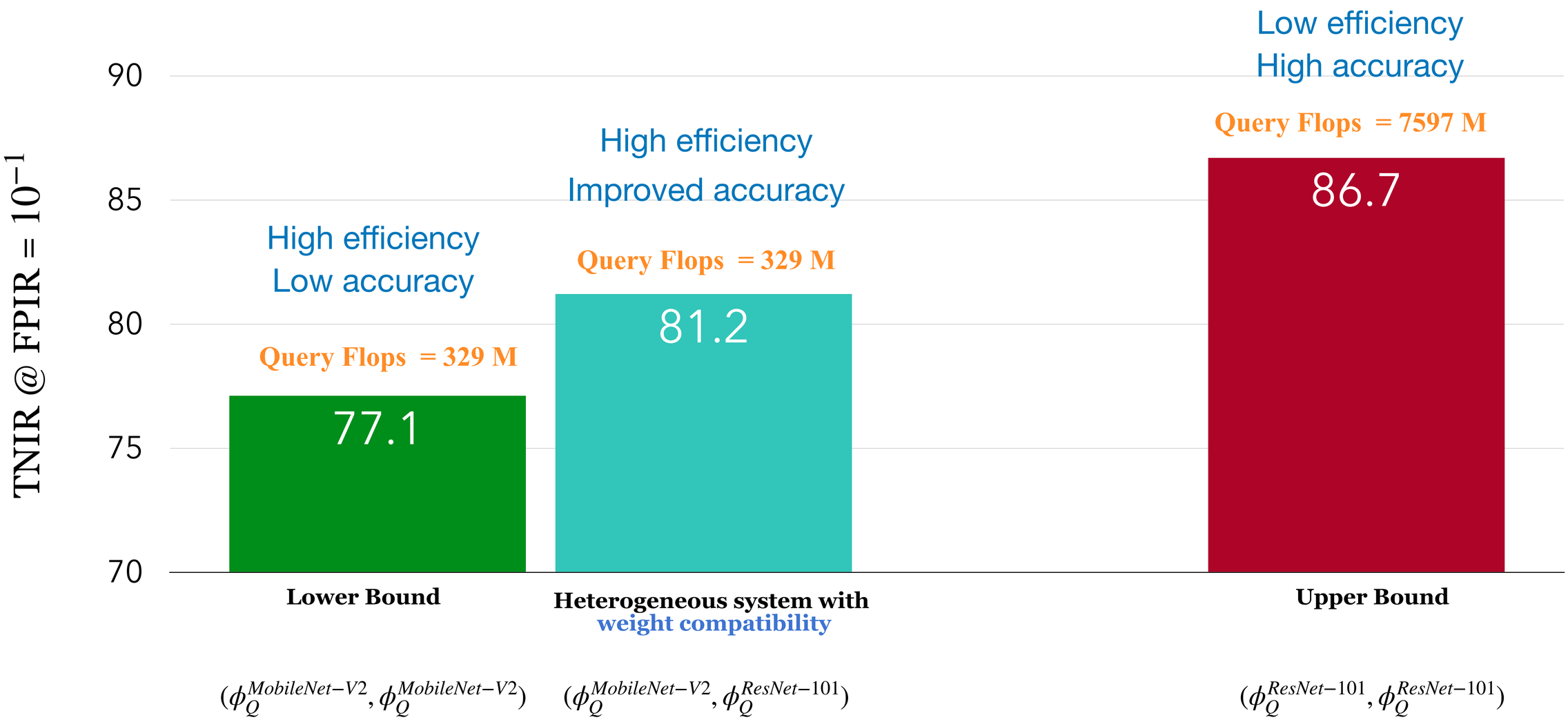
Results-1 (Face retrieval)



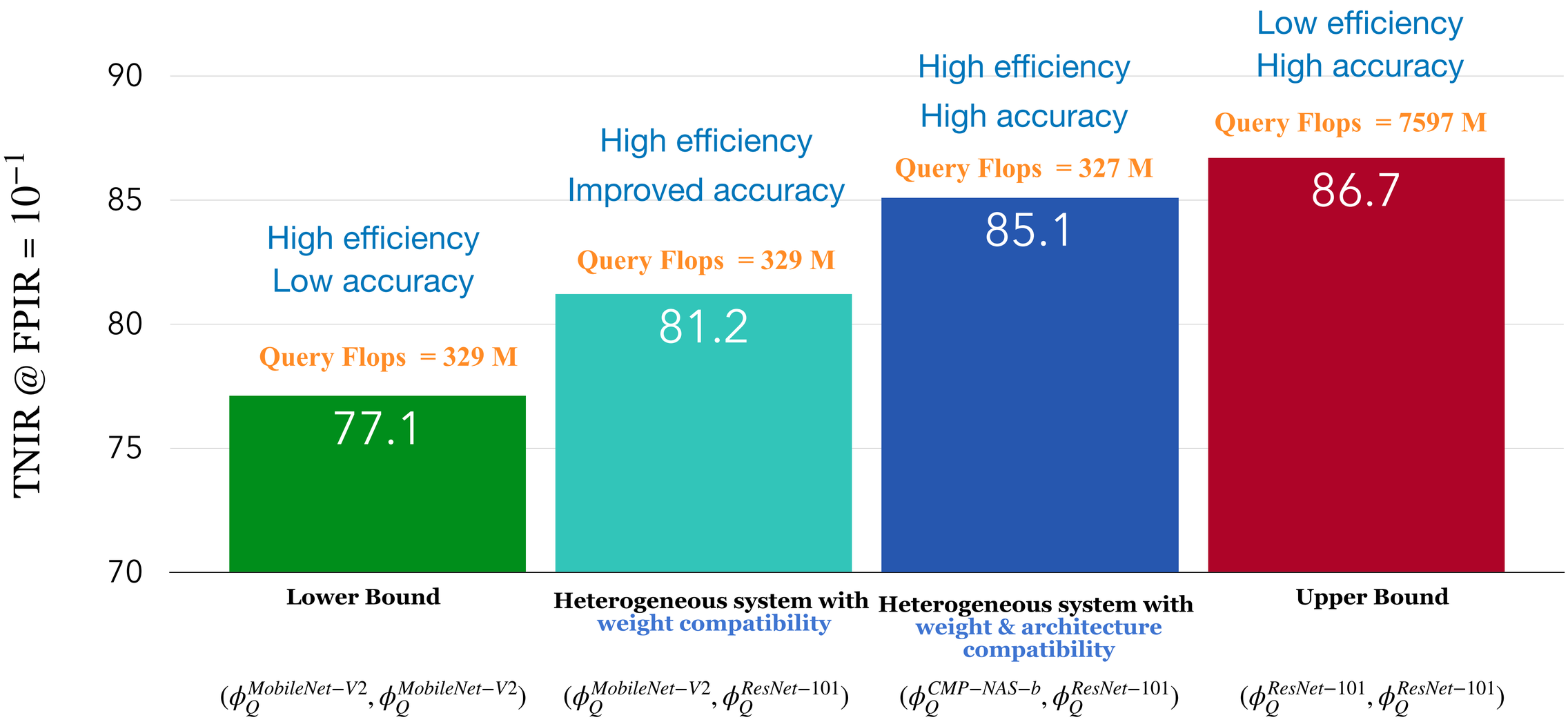
Results-1 (Face retrieval)



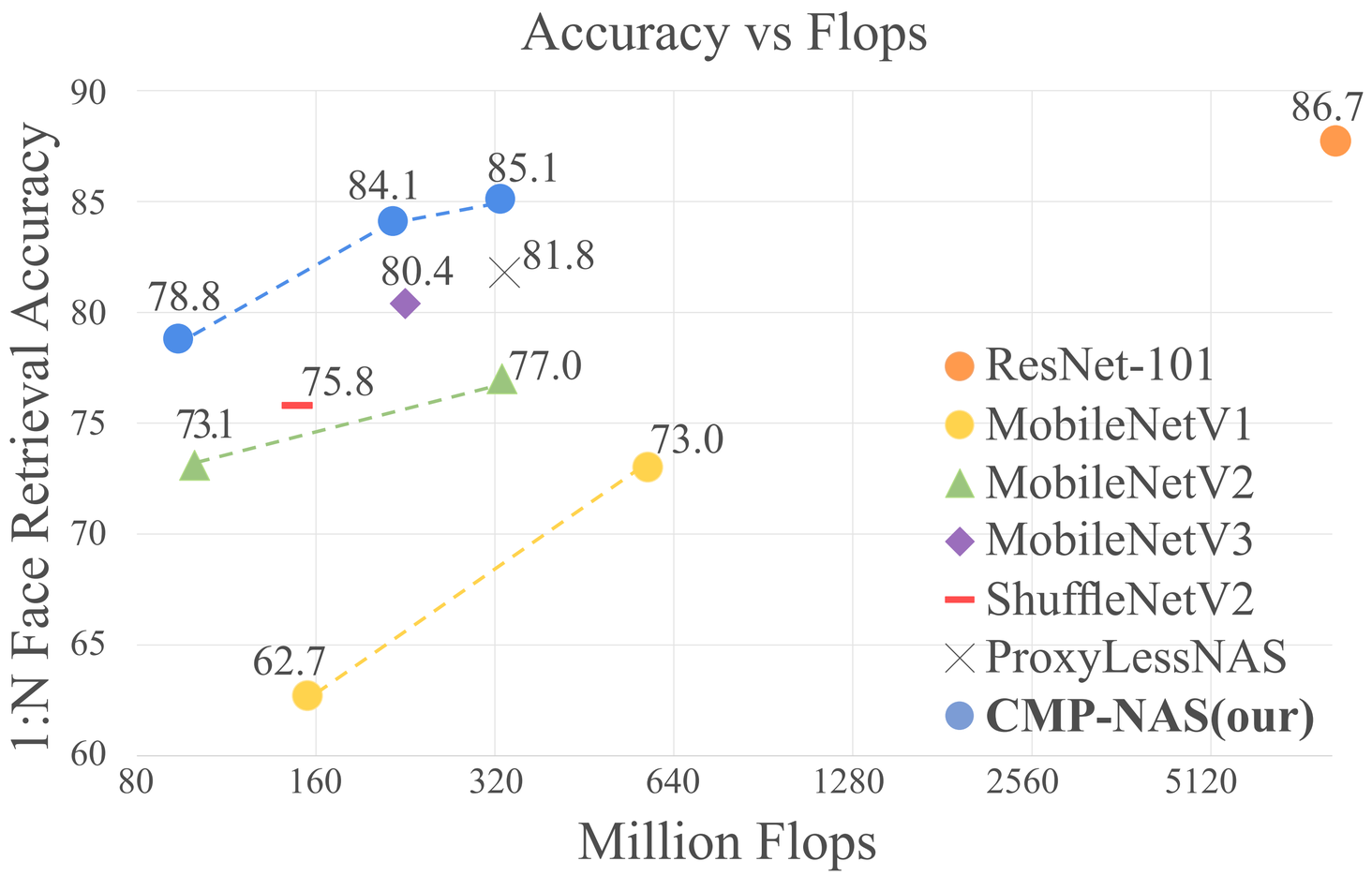
Results-1 (Face retrieval)



Results-1 (Face retrieval)

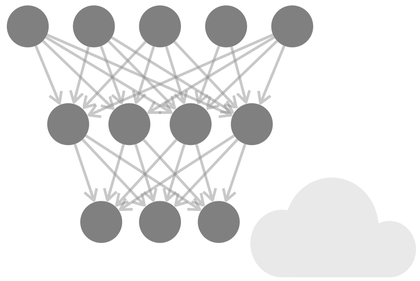


Ablations


Accuracy vs Flops
1:N Face Retrieval Accuracy
90
85
80
75
70
65
60
80
160
320
640
1280
2560
5120
Million Flops
86.7
84.1
85.1
81.8
80.4
78.8
75.8
77.0
73.1
73.0
62.7
ResNet-101
MobileNetV1
MobileNetV2
MobileNetV3
ShuffleNetV2
ProxyLessNAS
CMP-NAS(our)

Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

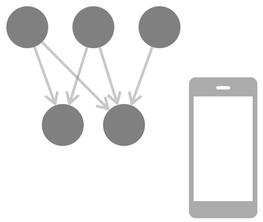
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

CUP: Cluster Pruning for Compressing Deep Neural Networks

Rahul Duggal, Cao Xiao, Richard Vuduc, Duen Horng Chau, Jimeng Sun

IEEE Big Data, 2021

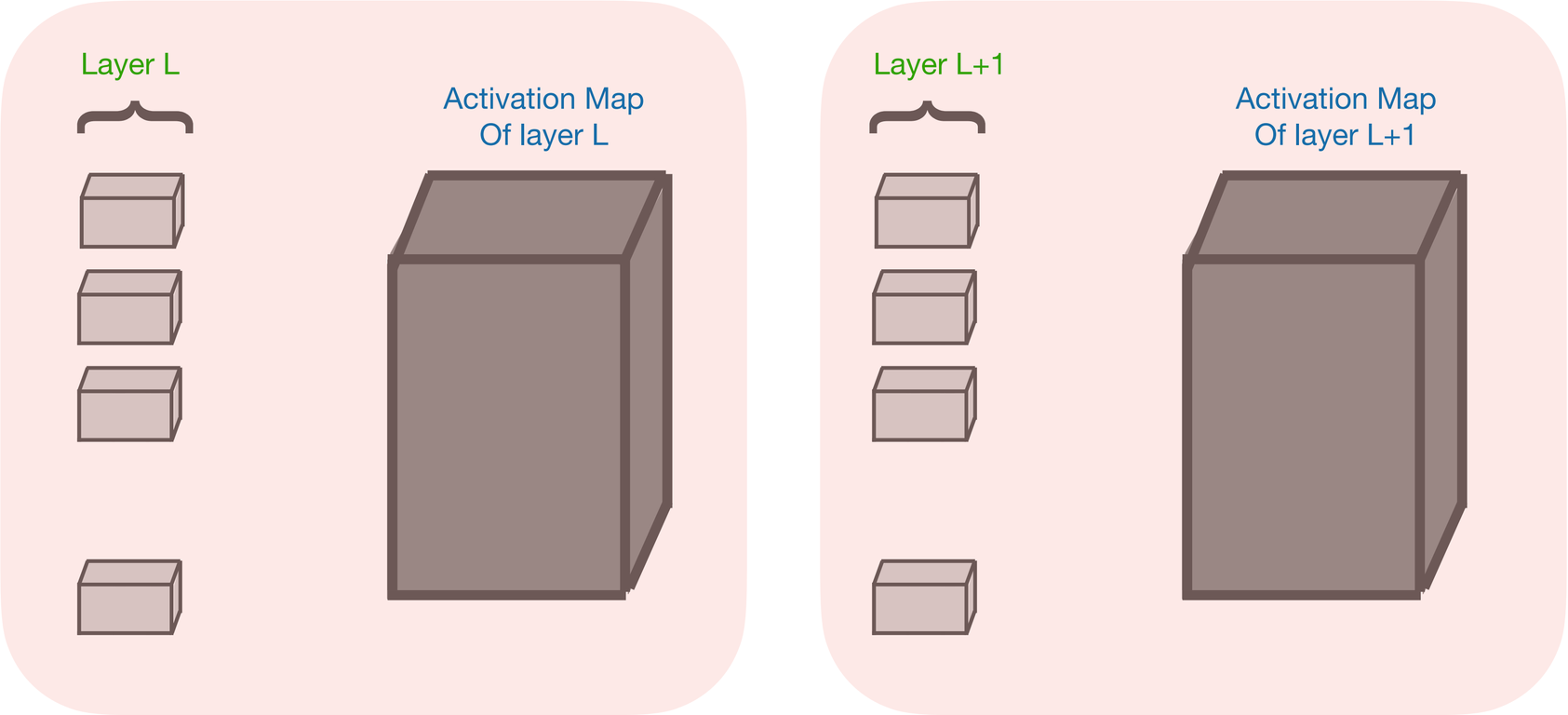
Goal

Reduce the **storage** and **computation** cost of a DNN

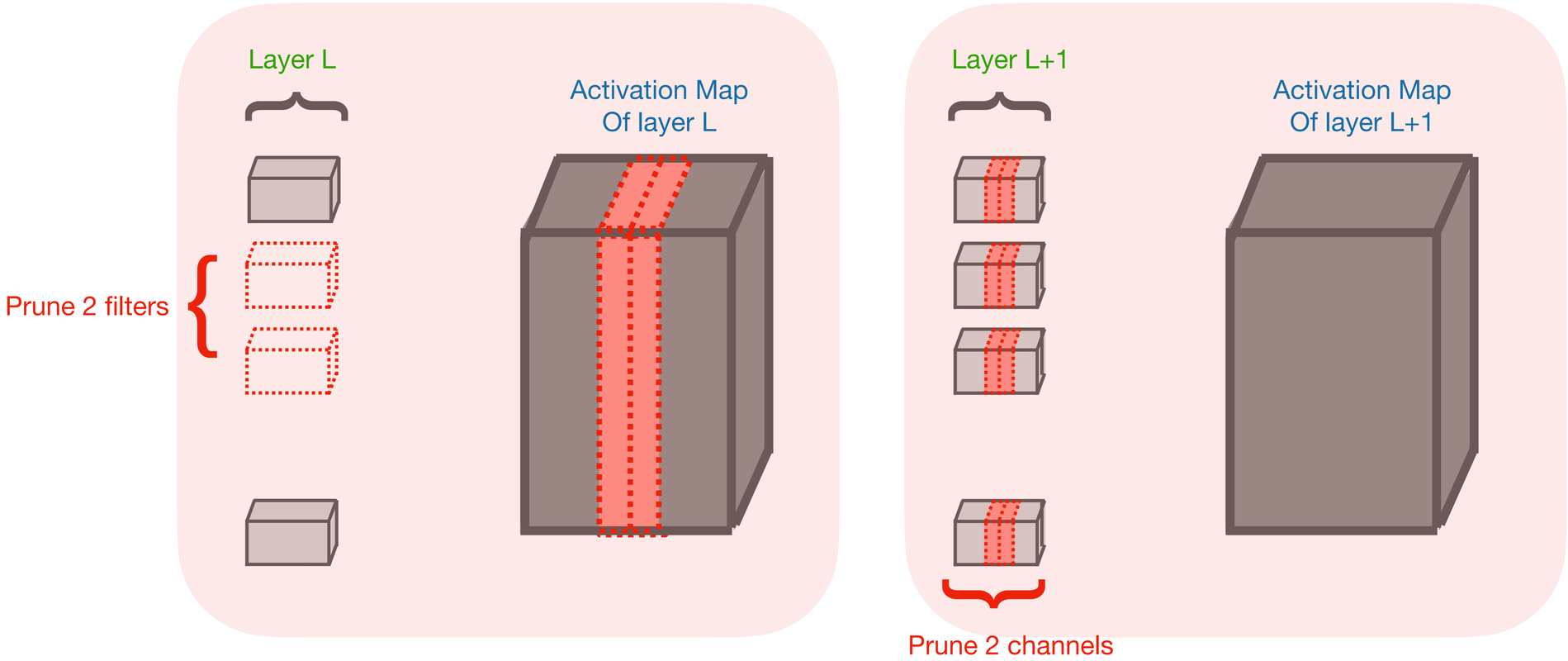
$$\mathbf{F}\,(x;\, W) \approx \mathbf{F}\,(x;\, W_{compressed})$$

Such that $|W_{compressed}| \ll |W|$

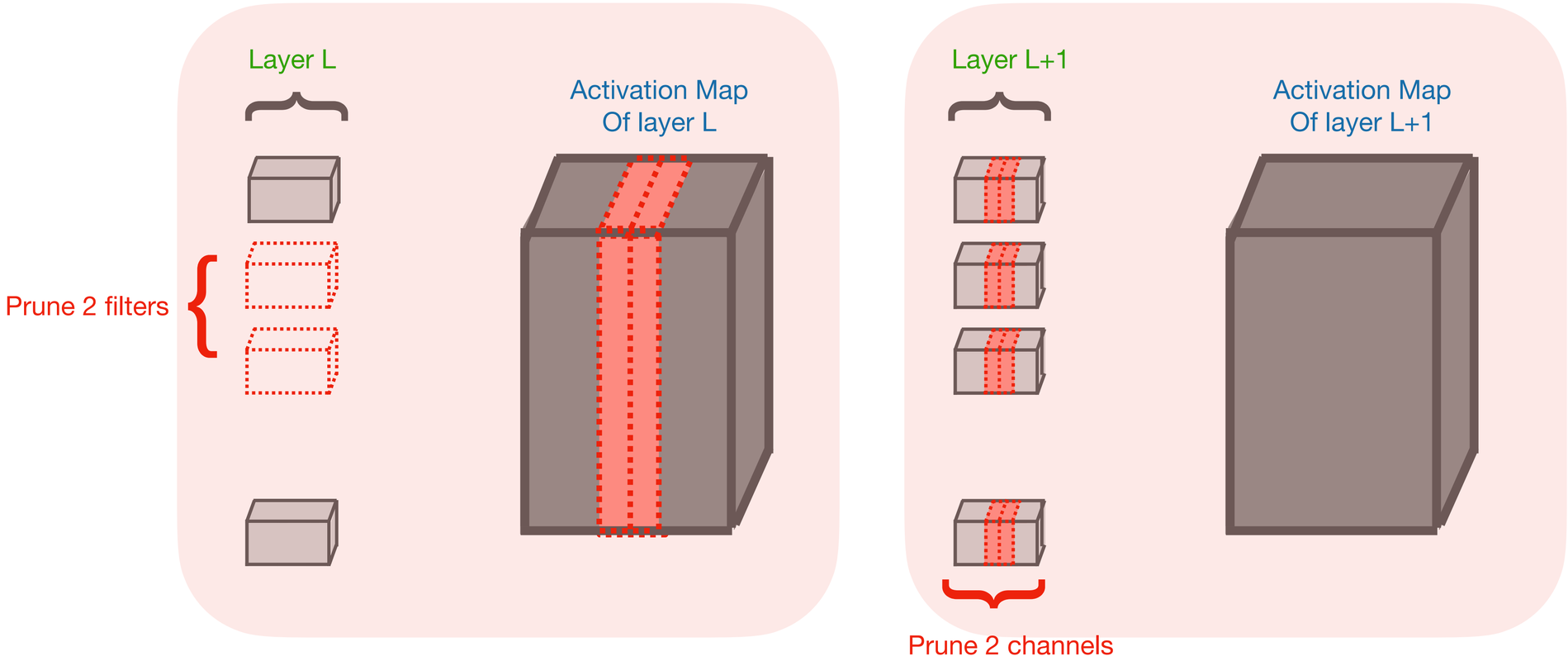
Filter Pruning



Filter Pruning



Filter Pruning



Which filters to prune?

Our method

CUP: Cluster Pruning

Our Idea: Prune **similar filters**

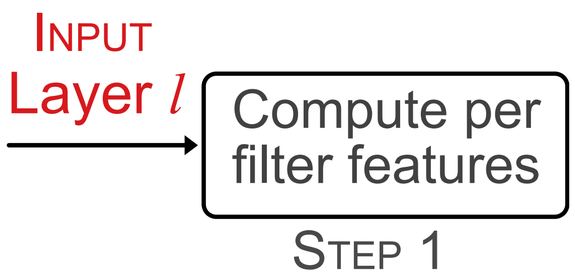
Our method

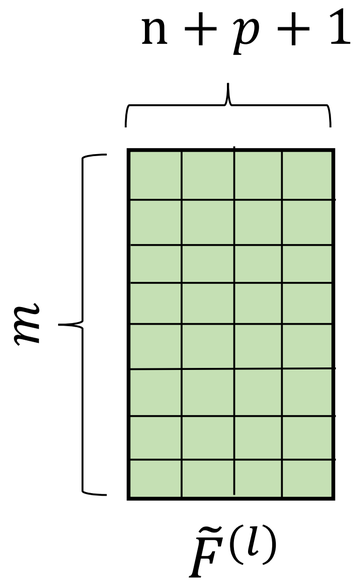
CUP: Cluster Pruning

Our Idea: Prune **similar filters**

Our method

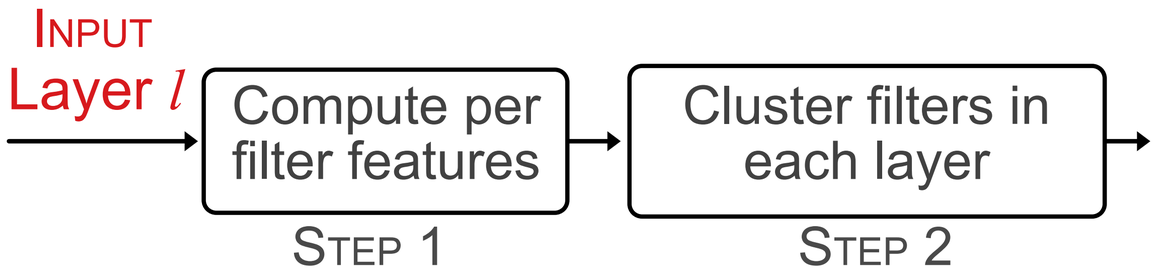
CUP: Cluster Pruning

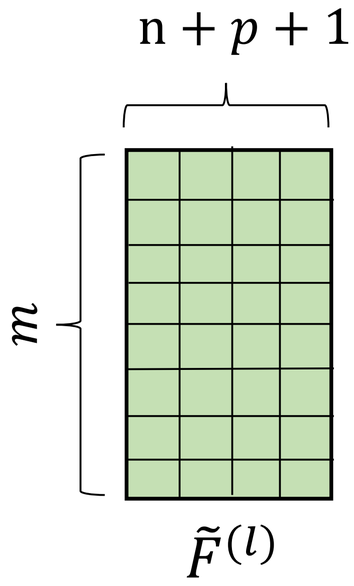




Our method

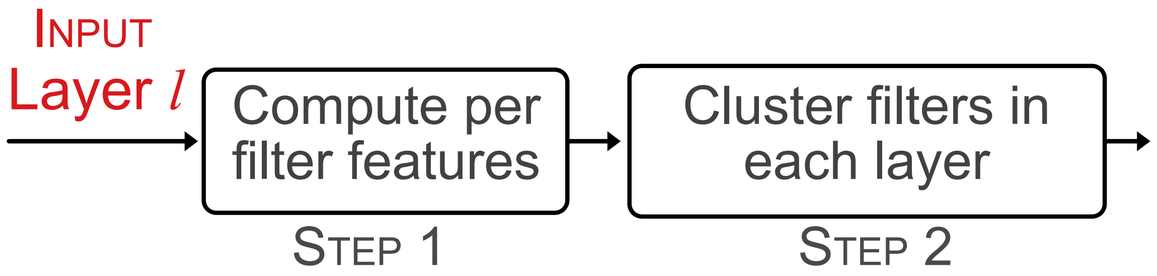
CUP: Cluster Pruning

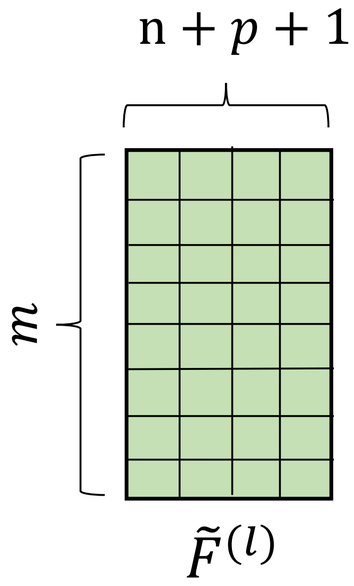




Our method

CUP: Cluster Pruning

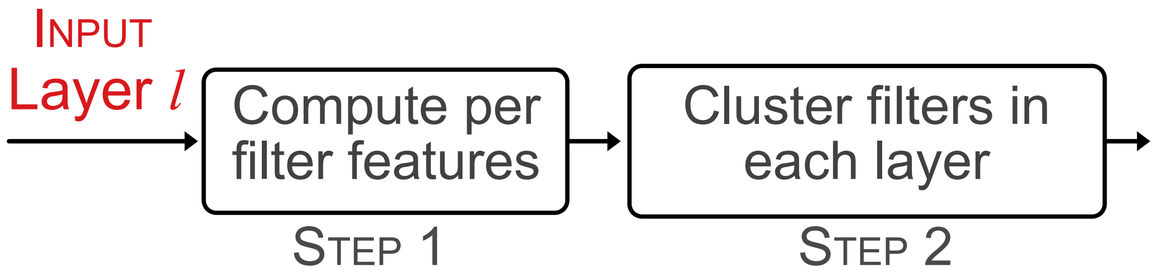


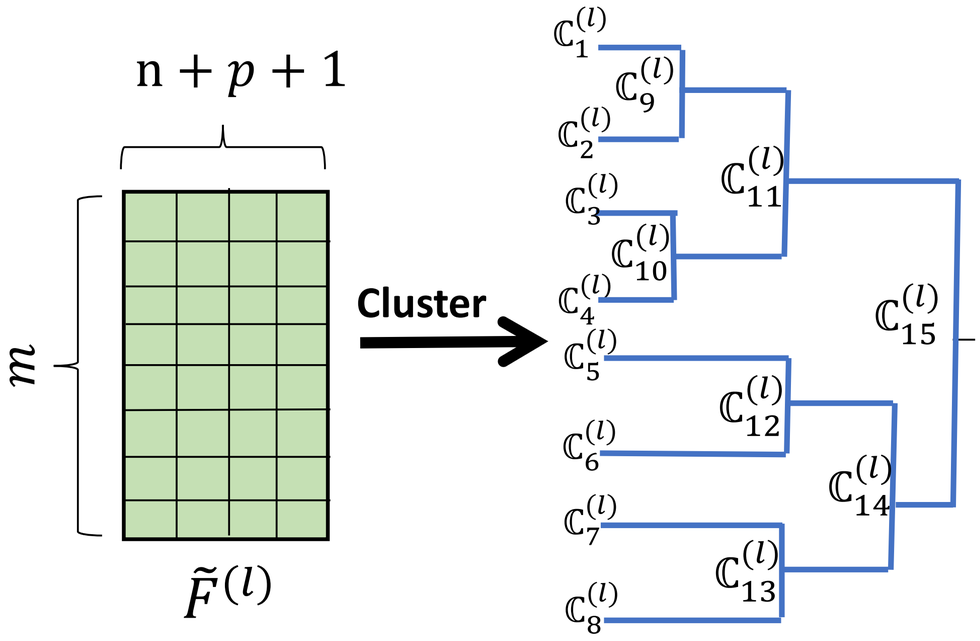


How many clusters?

Our method

CUP: Cluster Pruning

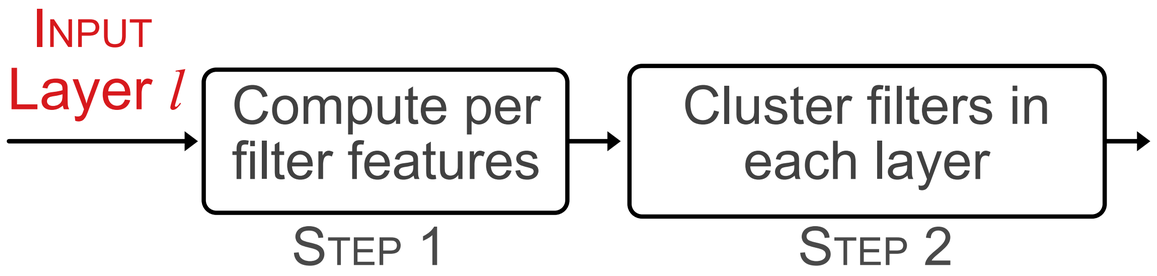


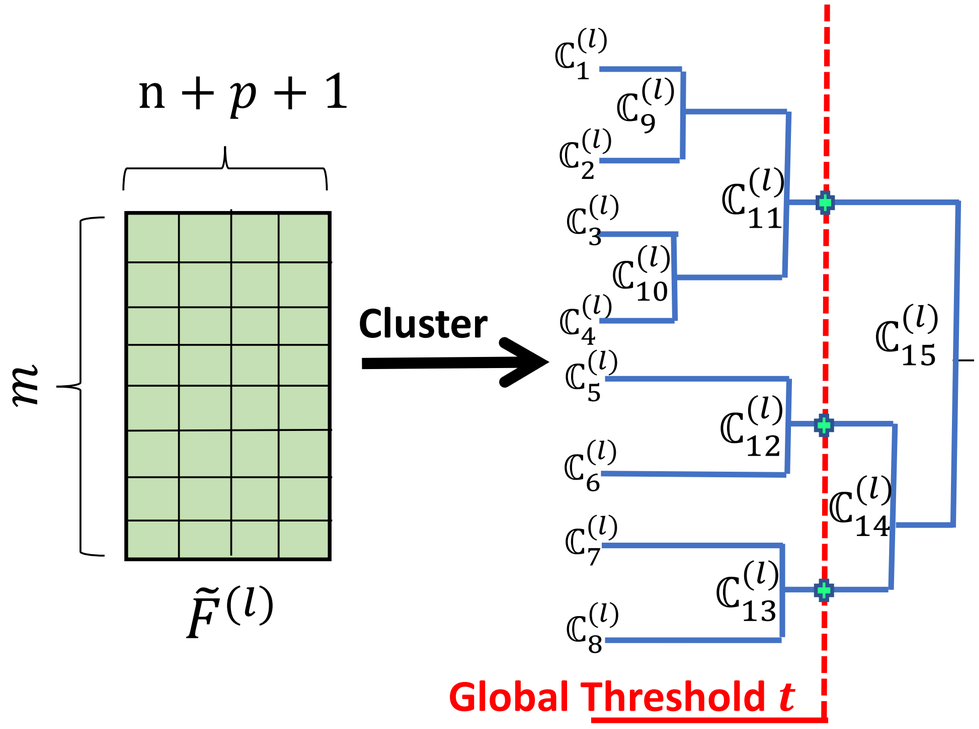


How many clusters?

Our method

CUP: Cluster Pruning

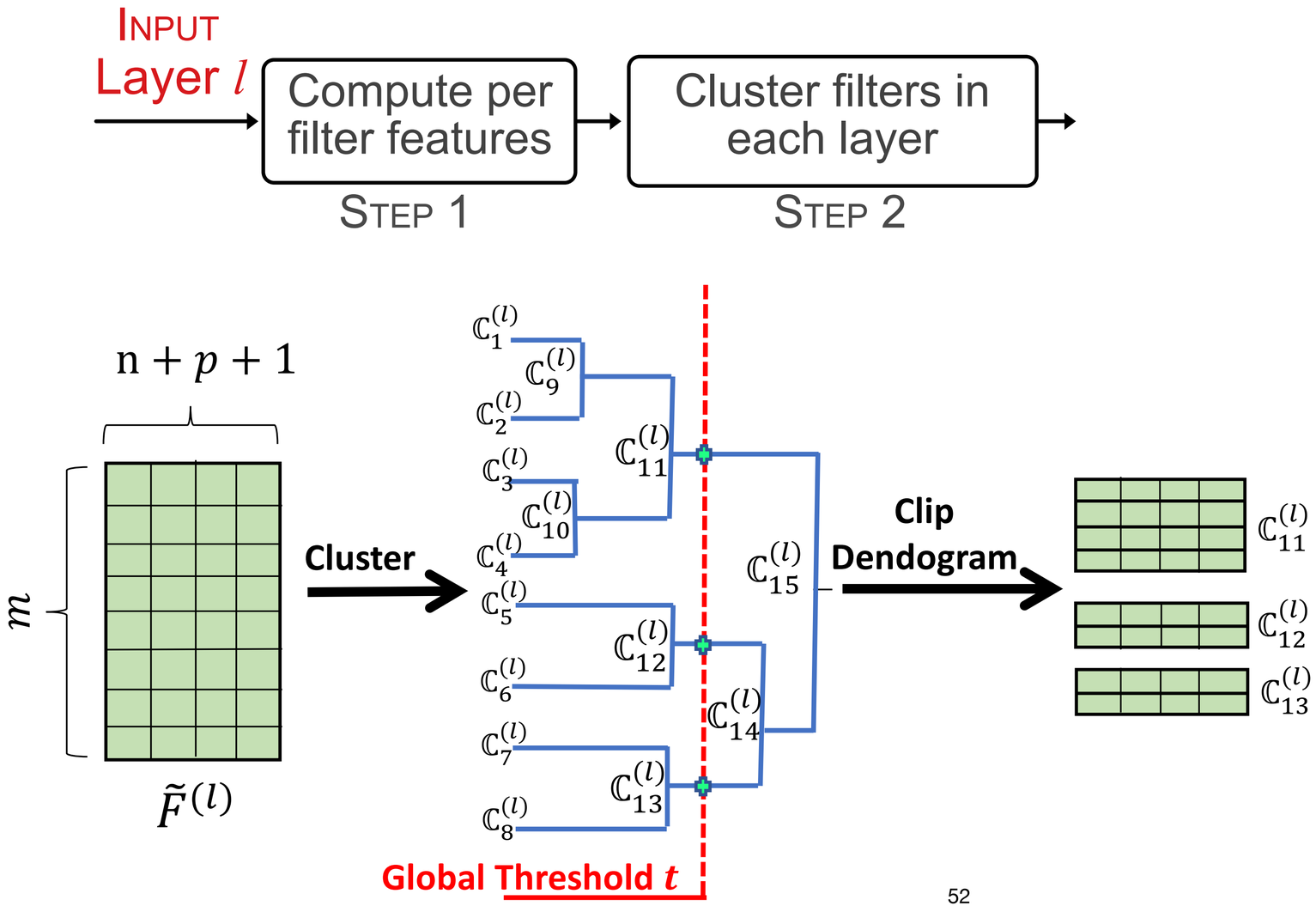




How many clusters?

Our method

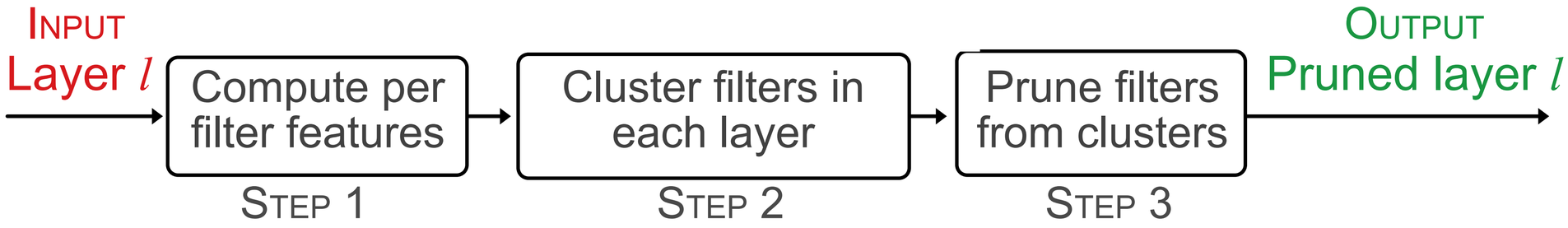
CUP: Cluster Pruning

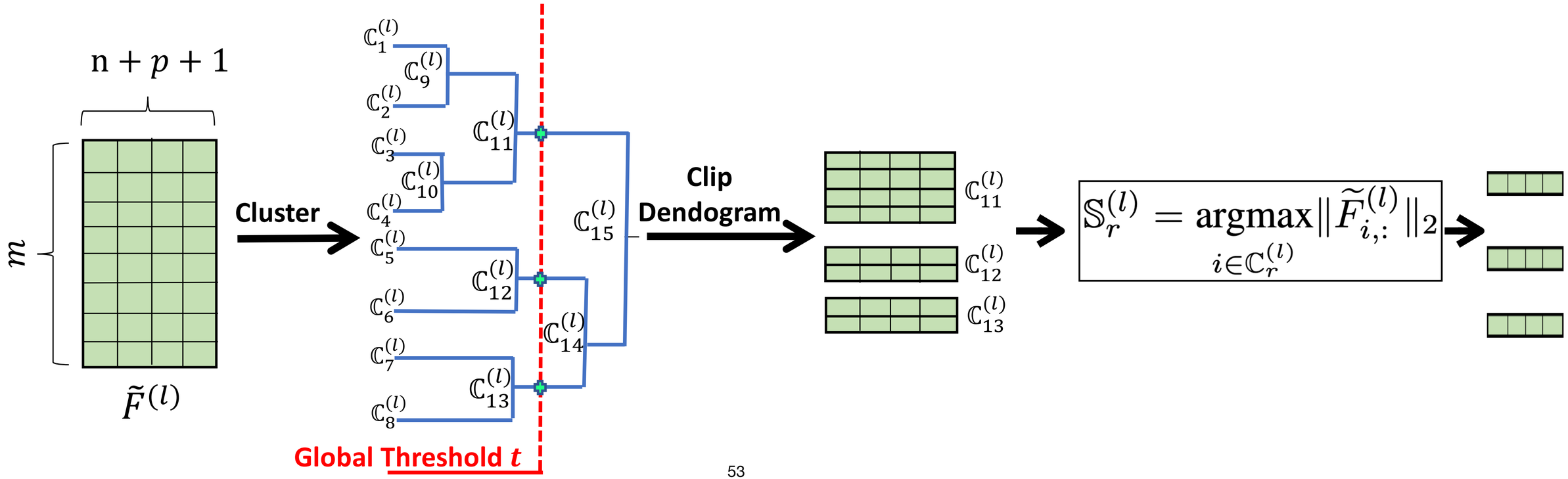


How many clusters?

Our method

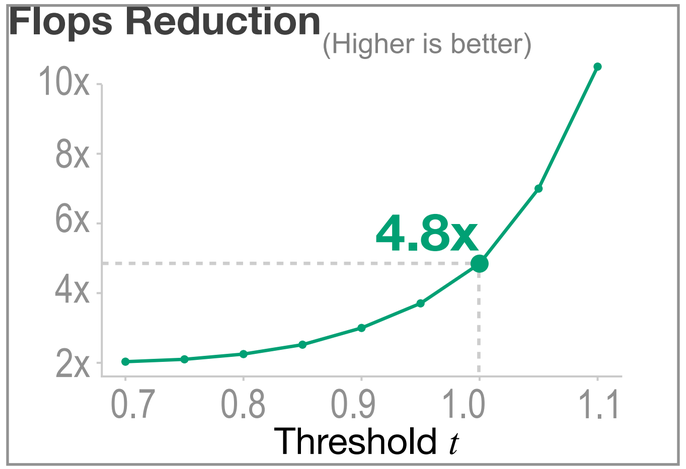
CUP: Cluster Pruning

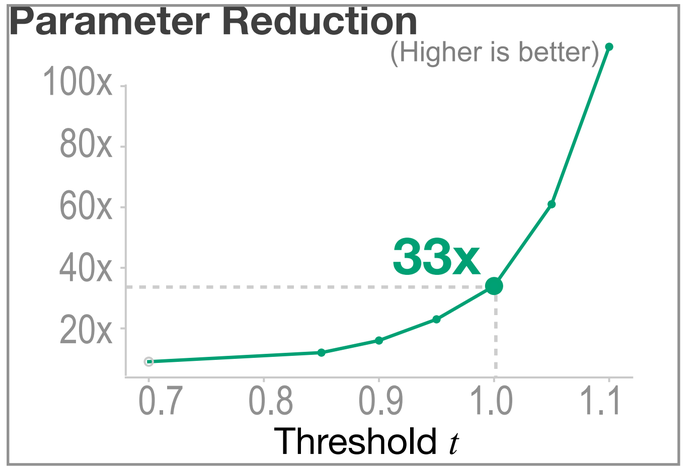


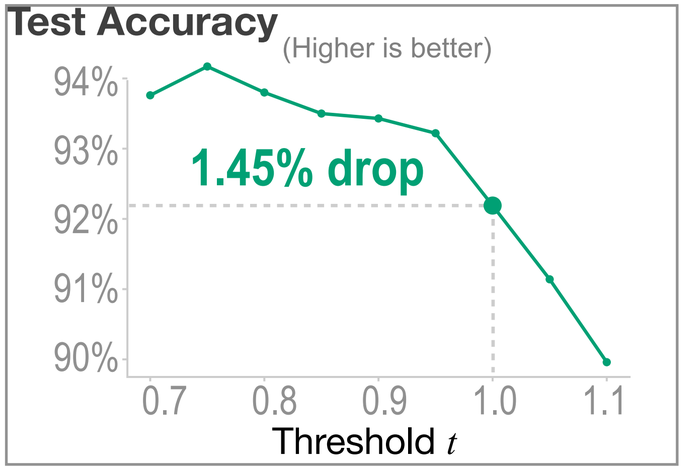


Results

Benefit 1: Single hyper parameter control over pruning amount







Results

Benefit 2: Training time reduction

Method	Retrain?	Top-1 (%)	FR (×)	Training Time (GPU Hours)
Resnet-50	-	75.86	1.00	66.0
SFP [14]	✗	74.01	1.73	61.8
GM [15]	✗	74.13	2.15	62.2
CUP-RF (ours)	✗	**74.34**	**2.21**	**51.6**

~15 hours saving with 2x compression

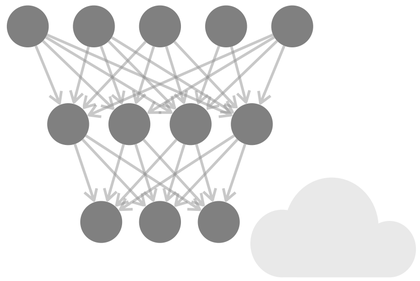
Results

Benefit 3: State-of-the-art compression

Model	Method	Retrain?	FR (×)	Acc. (Δ%)	
				Top-1	Top-5
ResNet-18	GM [15]	✓	1.71	-1.87	-1.15
	COP [29]	✓	**1.75**	-2.48	-
	CUP (Our)	✓	**1.75**	**-1.00**	**-0.79**
	SFP [14]	✗	1.71	-3.18	-1.85
	GM [15]	✗	1.71	-2.47	-1.52
	CUP-RF (ours)	✗	**1.75**	**-2.37**	**-1.40**
ResNet-34	L1 [2]	✓	1.31	-1.06	-
	GM [15]	✓	1.69	-1.29	-0.54
	CUP (ours)	✓	**1.78**	**-0.86**	**-0.53**
	SFP [14]	✗	1.69	-2.09	-1.29
	GM [15]	✗	1.69	-2.13	-0.92
	CUP-RF (ours)	✗	**1.71**	**-1.61**	**-0.89**
ResNet-50	SFP [14]	✓	2.15	-14.0	-8.20
	MP [30]	✓	2.05	-1.20	-
	CUP (ours)	✓	**2.47**	**-1.17**	**-0.81**
	SFP [14]	✗	1.71	-1.54	**-0.81**
	GM [15]	✗	2.15	-2.02	-0.93
	CUP-RF (ours)	✗	**2.20**	**-1.47**	-0.88

Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

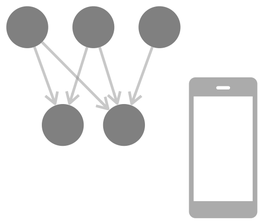
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

REST: Robust and Efficient Sleep Staging in the Wild

Rahul Duggal*, Scott Freitas*, Cao Xiao, Duen Horng Chau, Jimeng Sun

The World Wide Web Conference (**WWW ’20**)

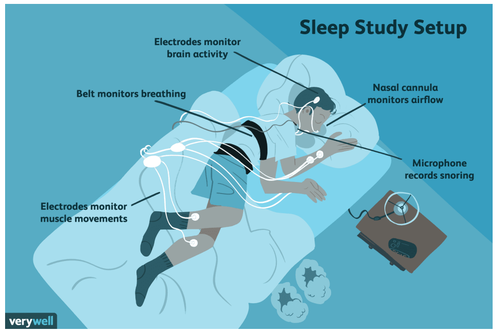
Why sleep diagnosis matters?

Huge cost of lost productivity!

Hafner, Marco et al. “Why Sleep Matters-The Economic Costs of Insufficient Sleep: A Cross-Country Comparative Analysis.” Rand health quarterly vol. 6,4 11. 1 Jan. 2017

Current workflow of sleep diagnosis

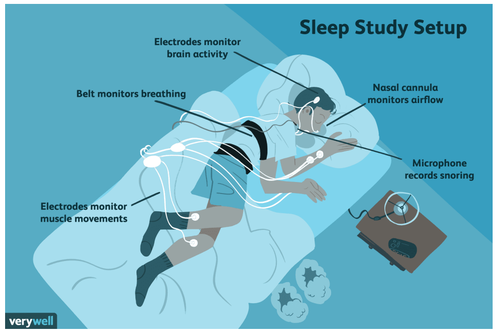
1. Overnight Sleep Exam



https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

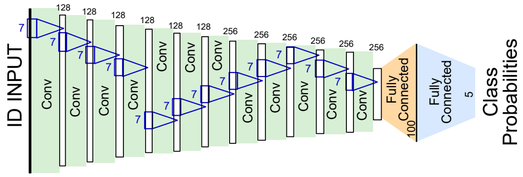
Current workflow of sleep diagnosis

1. Overnight Sleep Exam



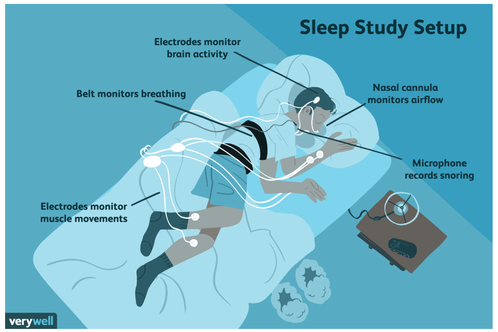
https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

2. Deep neural network



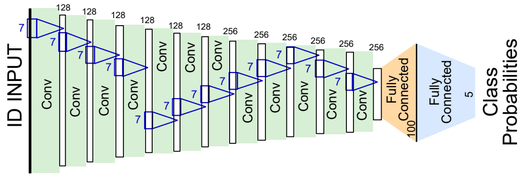
Current workflow of sleep diagnosis

1. Overnight Sleep Exam

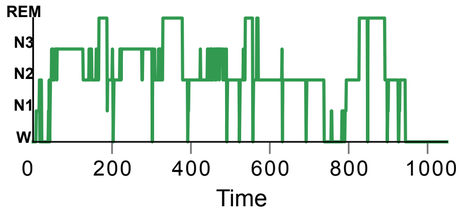


https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

2. Deep neural network

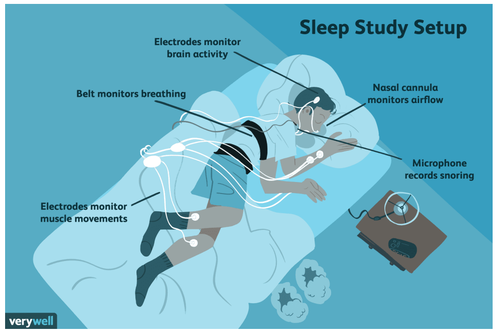


3. Hypnogram



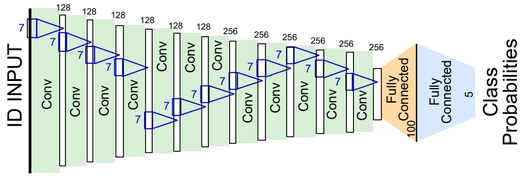
Current workflow of sleep diagnosis

1. Overnight Sleep Exam

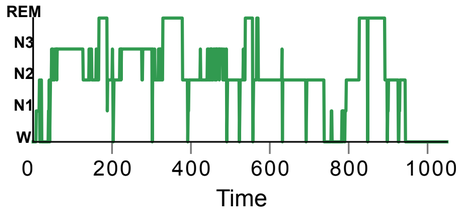


https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

2. Deep neural network



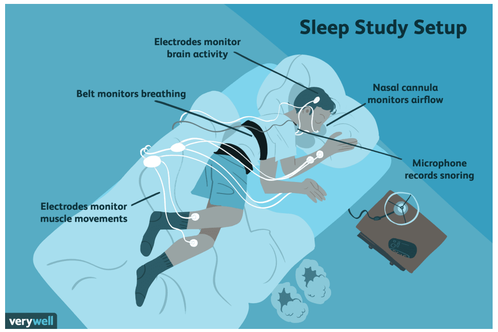
3. Hypnogram



4. Doctor's analyze

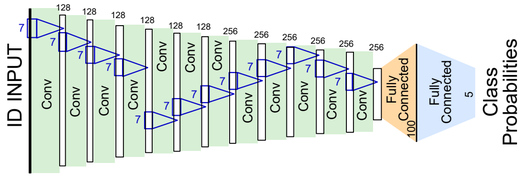
Current workflow of sleep diagnosis

1. Overnight Sleep Exam

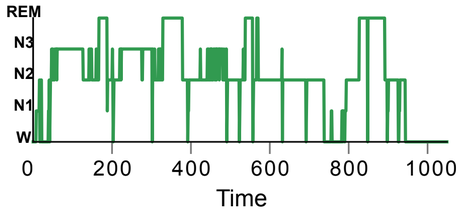


https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

2. Deep neural network



3. Hypnogram

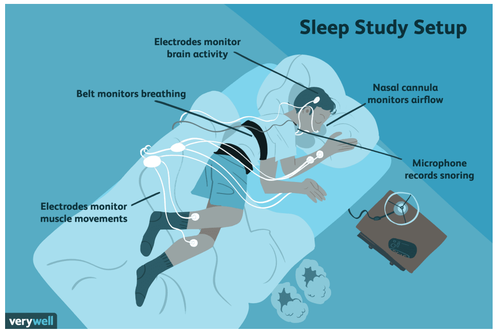


4. Doctor's analyze

Diagnosis

Current workflow of sleep diagnosis

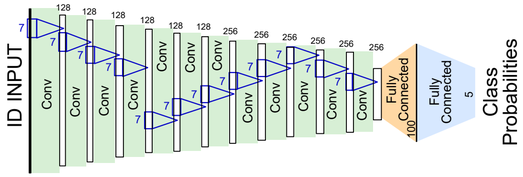
1. Overnight Sleep Exam



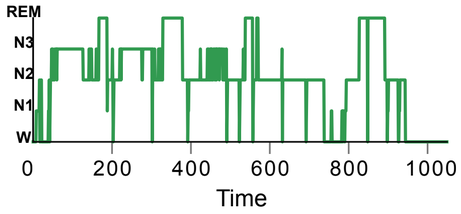
https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

1. Costly
2. Invasive
3. Inconvenient

2. Deep neural network



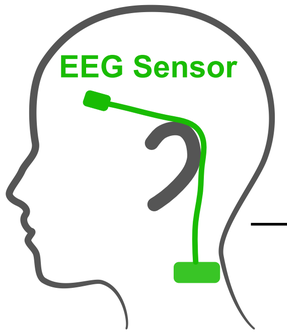
3. Hypnogram



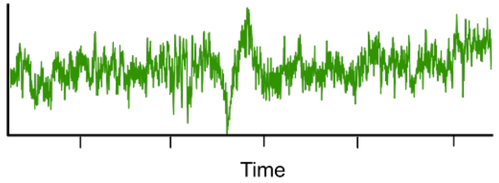
4. Doctor's analyze

Diagnosis

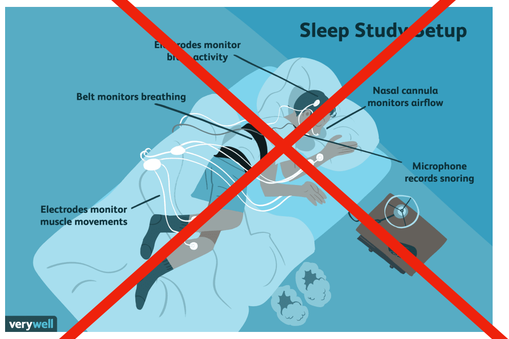
Current workflow of sleep diagnosis



Wearable based sleep monitoring



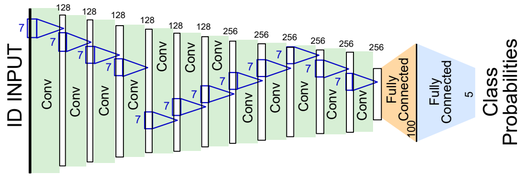
1. Overnight Sleep Exam



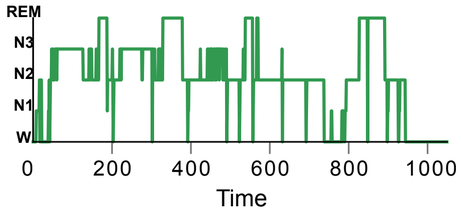
https://www.verywellhealth.com/what-to-expect-in-a-sleep-study-3015121

1. Costly
2. Invasive
3. Inconvenient

2. Deep neural network



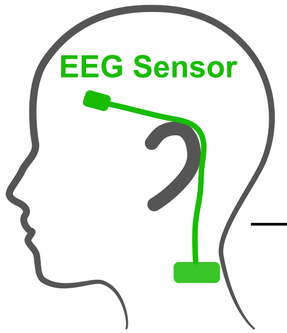
3. Hypnogram



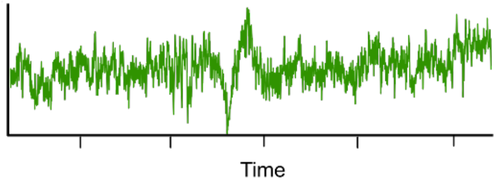
4. Doctor's analyze

Diagnosis

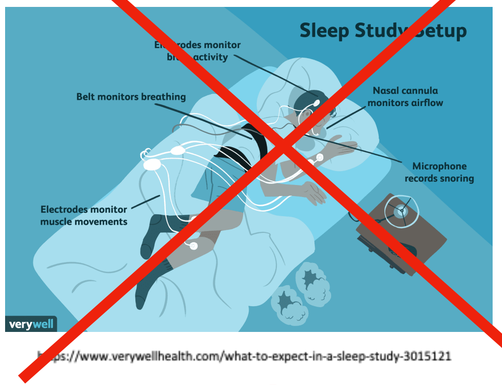
Current workflow of sleep diagnosis



Wearable based sleep monitoring

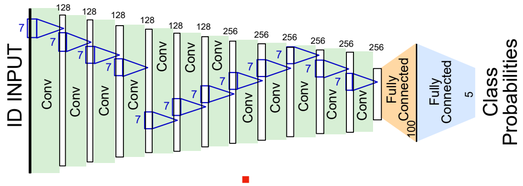


1. Overnight Sleep Exam



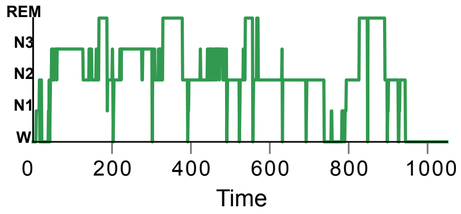
1. Costly
2. Invasive
3. Inconvenient

2. Deep neural network



1. Susceptible to noise.
2. Large compute requirement.

3. Hypnogram

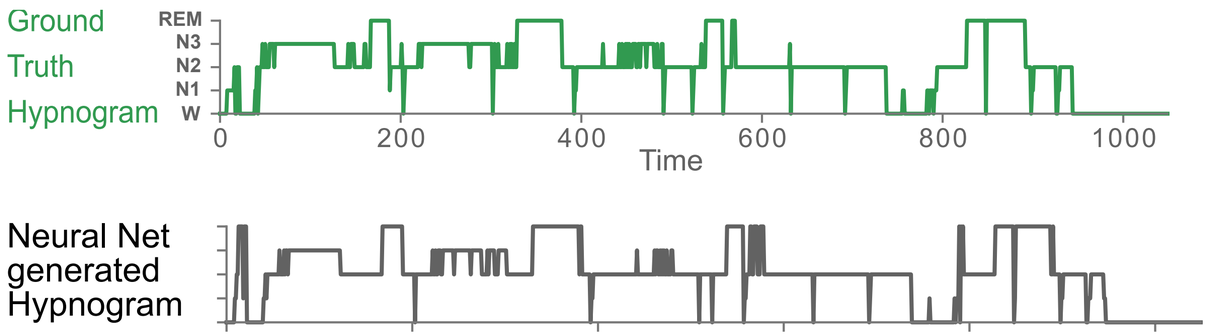


4. Doctor's analyze

Diagnosis

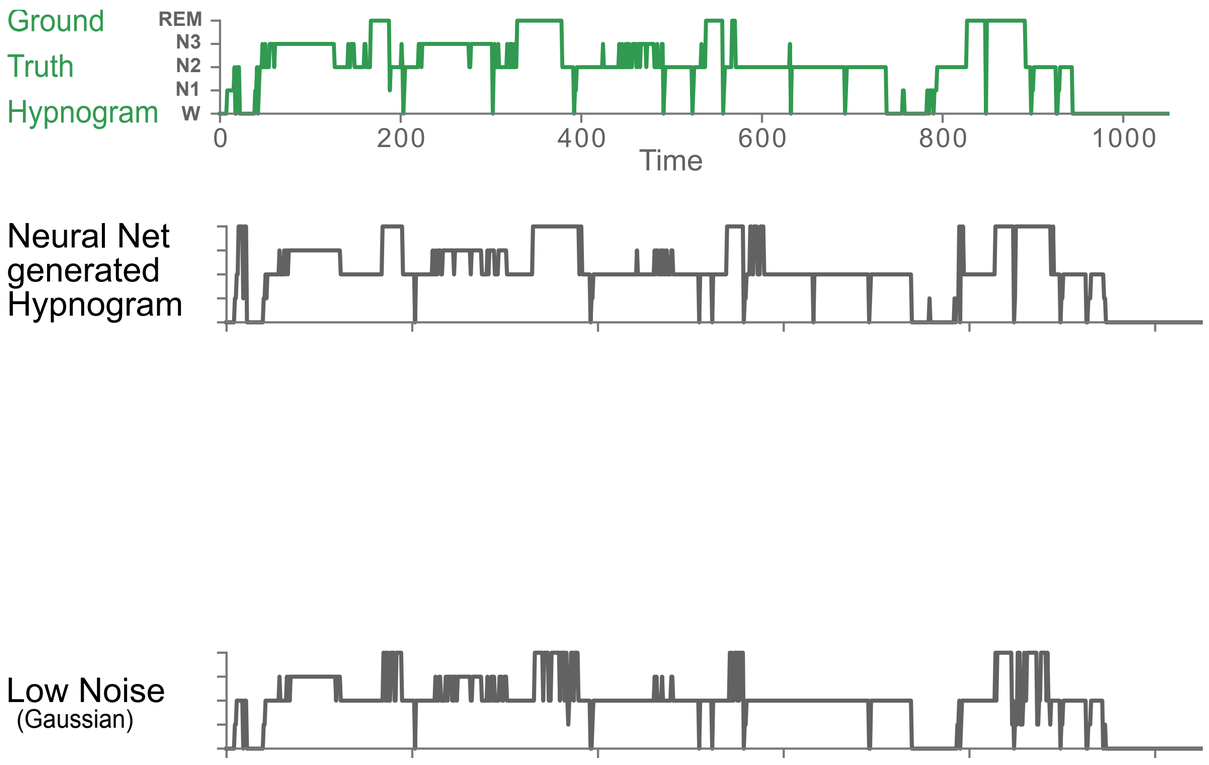
Research challenge 1

Robustness to Noise



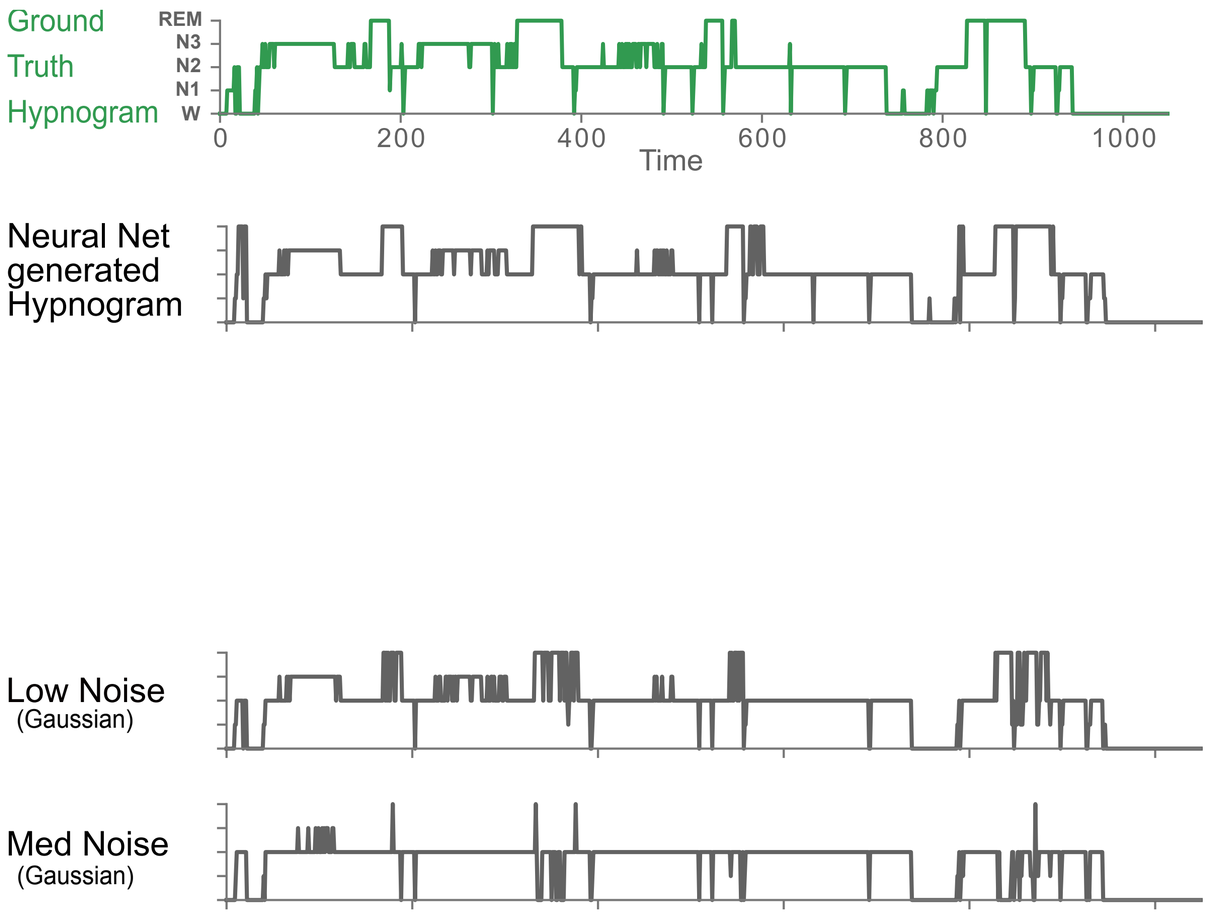
Research challenge 1

Robustness to Noise



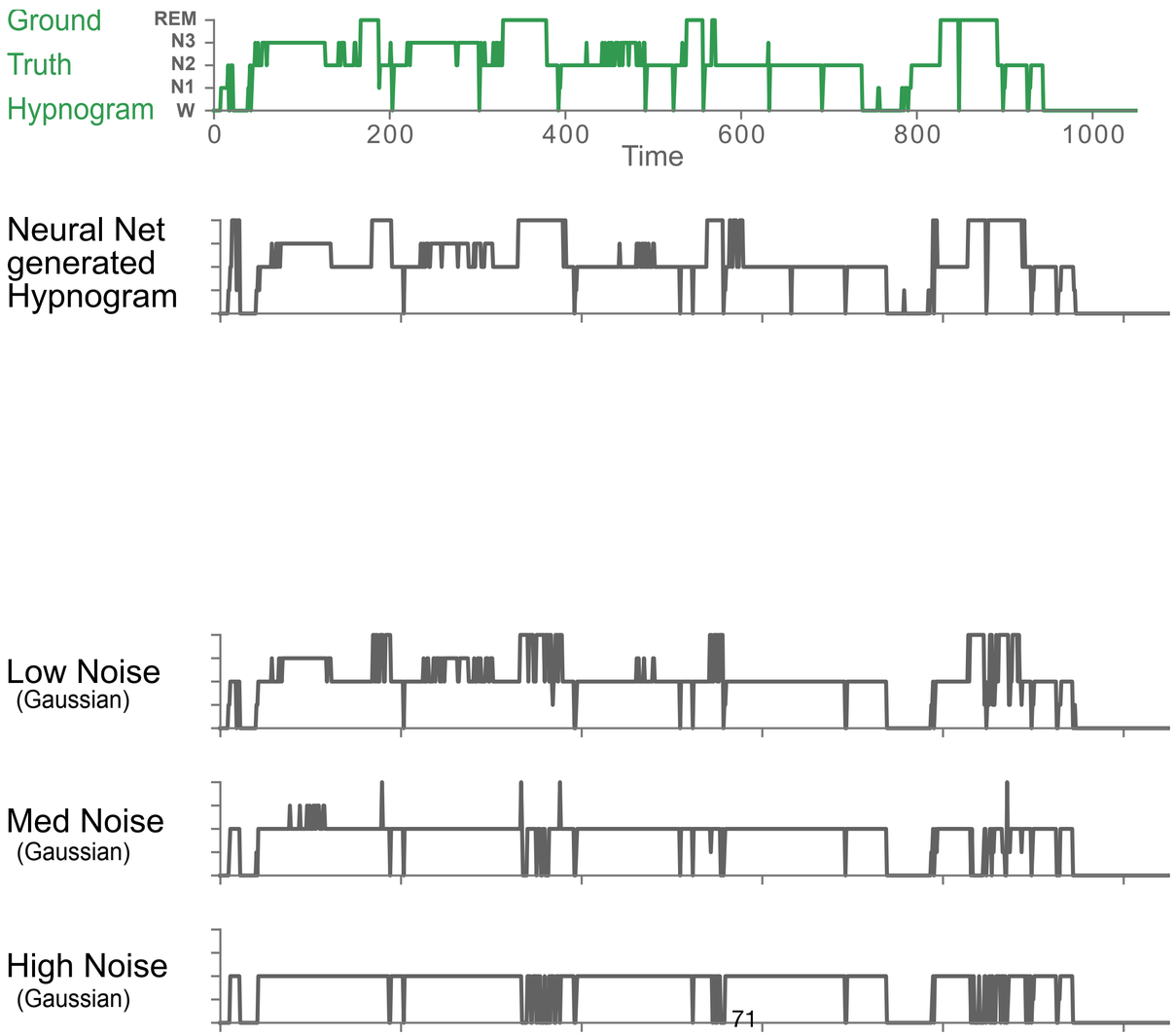
Research challenge 1

Robustness to Noise



Research challenge 1

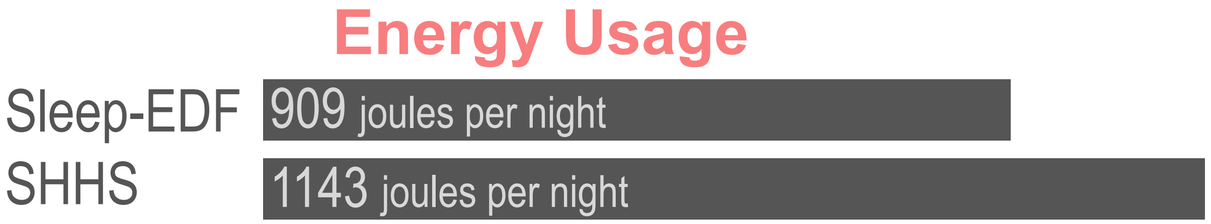
Robustness to Noise



Research challenge 2

Large compute requirements

>5 minutes!



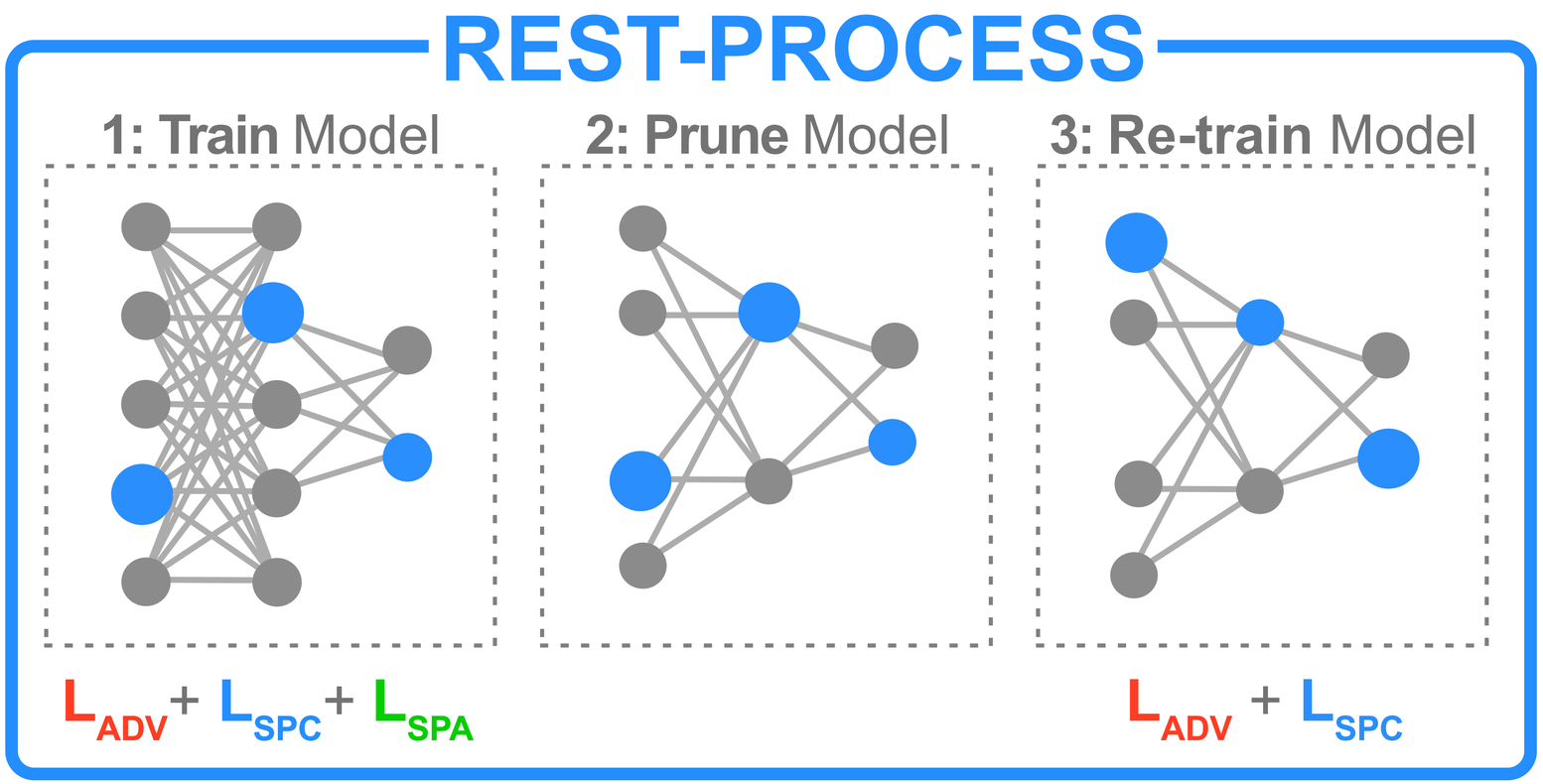
Fast battery draining!

REST: Our Goals

Develop neural networks for home sleep monitoring that are:

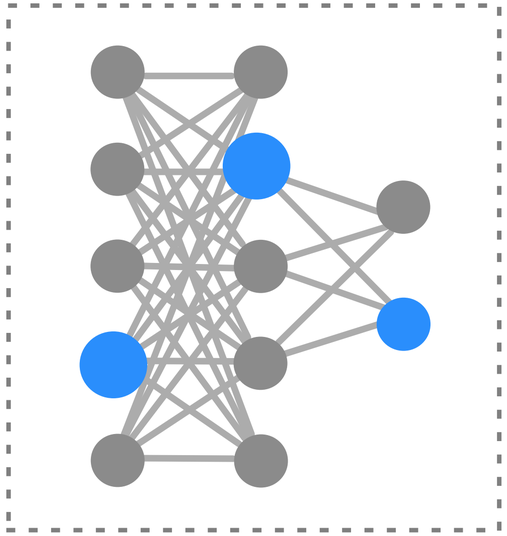
1. Robust to Noise
2. Energy and Compute Efficient

Method


REST-PROCESS
1: Train Model
2: Prune Model
3: Re-train Model
$L_{ADV} + L_{SPC} + L_{SPA}$
$L_{ADV} + L_{SPC}$

Method

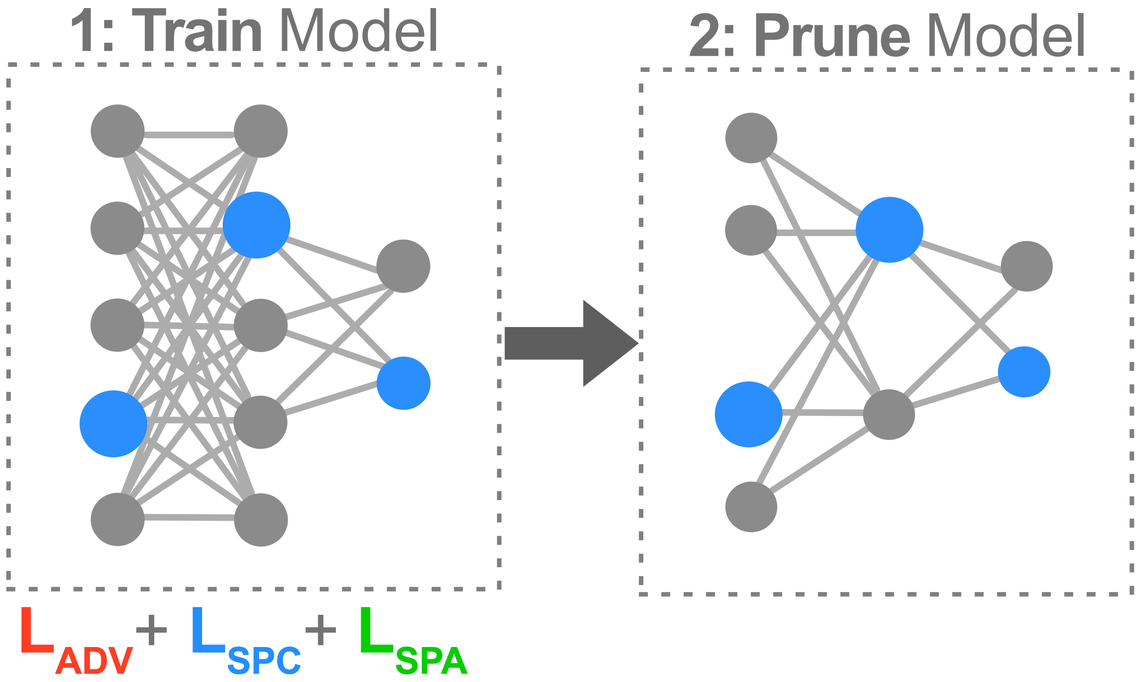
1: Train Model



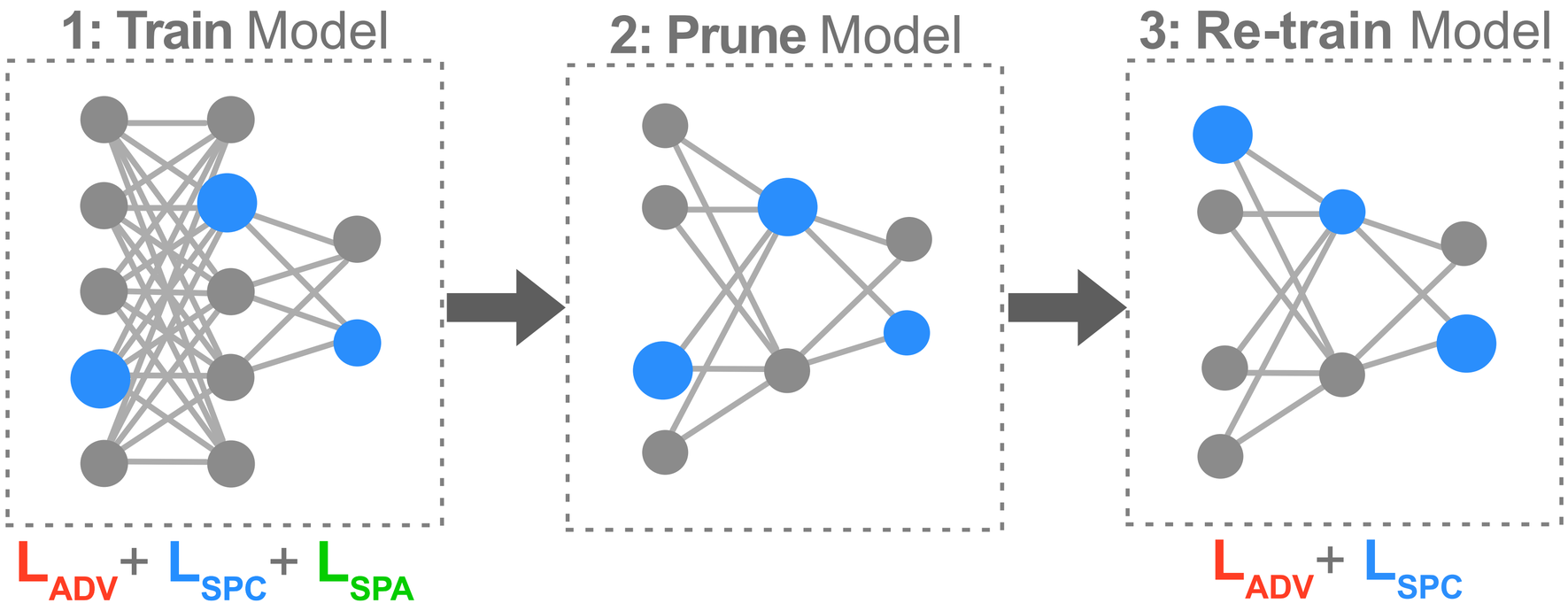
$L_{ADV} + L_{SPC} + L_{SPA}$

Loss Term	Definition	Intuition
L_{ADV}	$L(f(X^{n_{iter}}; W), y)$	Enable model to get used to noisy input.
L_{SPC}	$\sum_{i=1}^{N} \|(W^{(i)})^T W^{(i)} - I\|$	Encourage model parameters to be less sensitive to noise (small input changes → small out changes)
L_{SPA}	$\sum_{i=1}^{N} \|\gamma^{(l)}\|$	Enables filter importance estimation

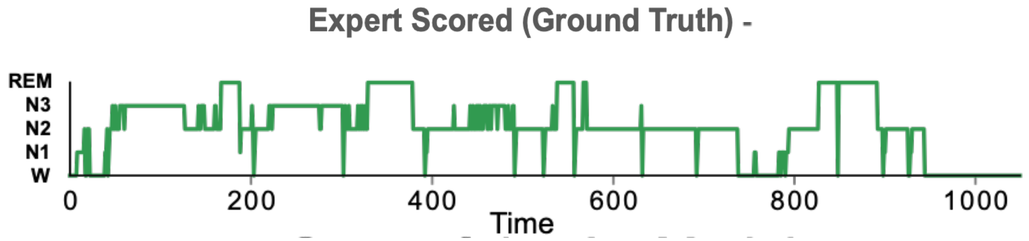
Method

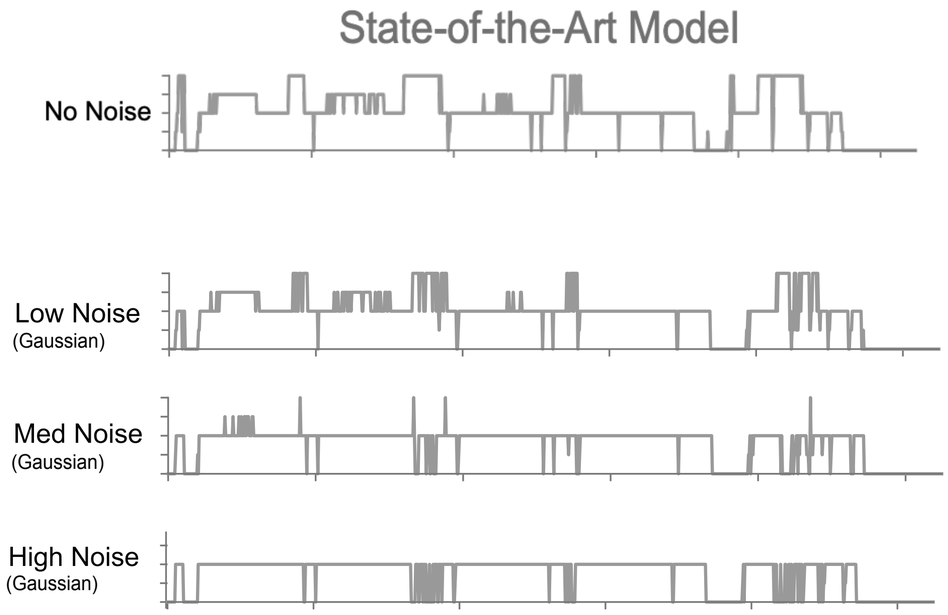


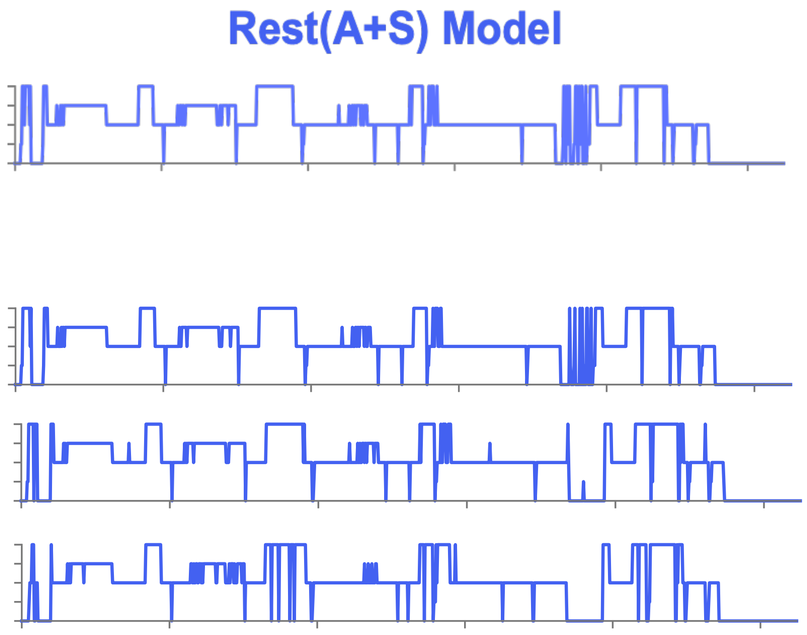
Method


1: Train Model
2: Prune Model
3: Re-train Model
$L_{ADV} + L_{SPC} + L_{SPA}$
$L_{ADV} + L_{SPC}$

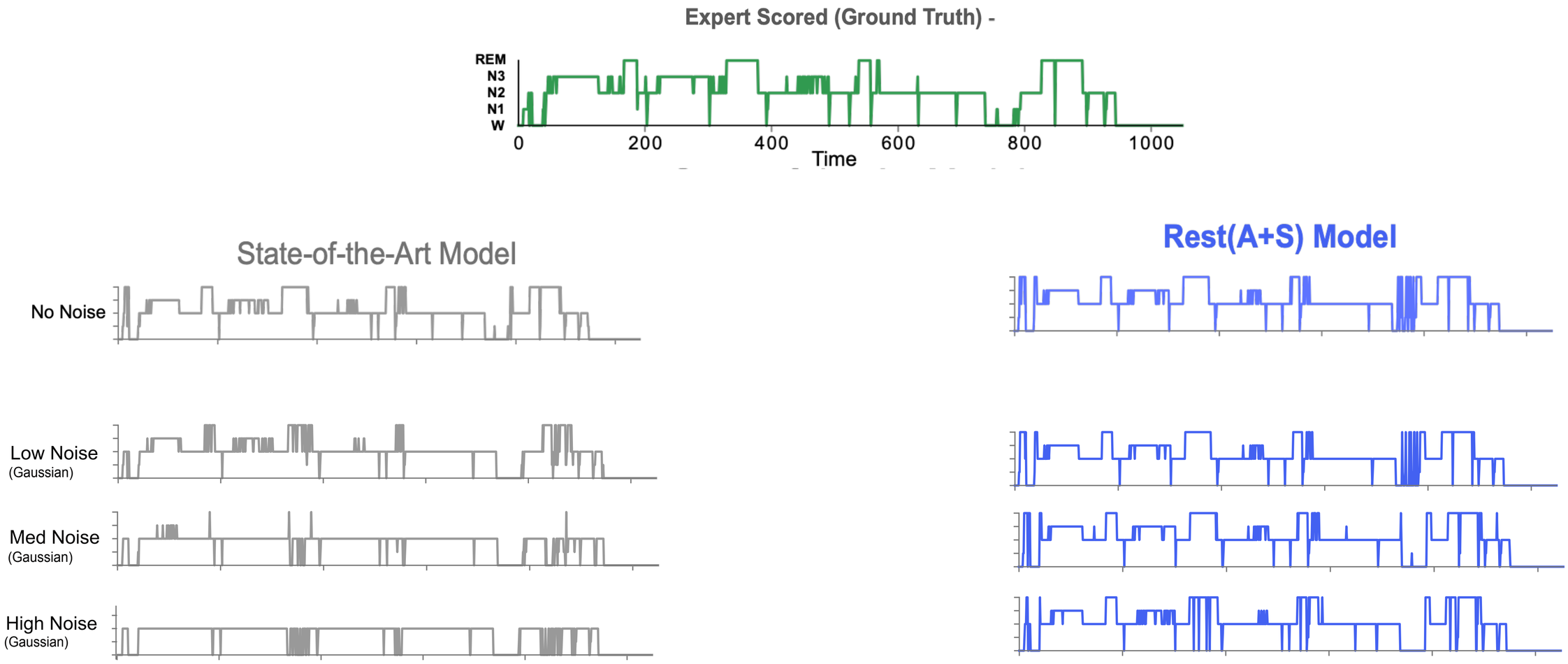
Results: Robustness to Noise



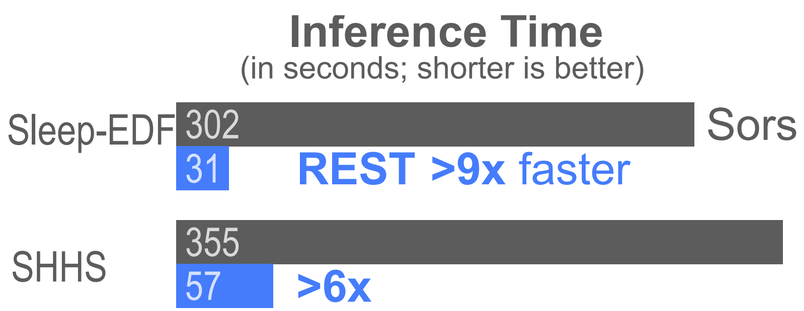


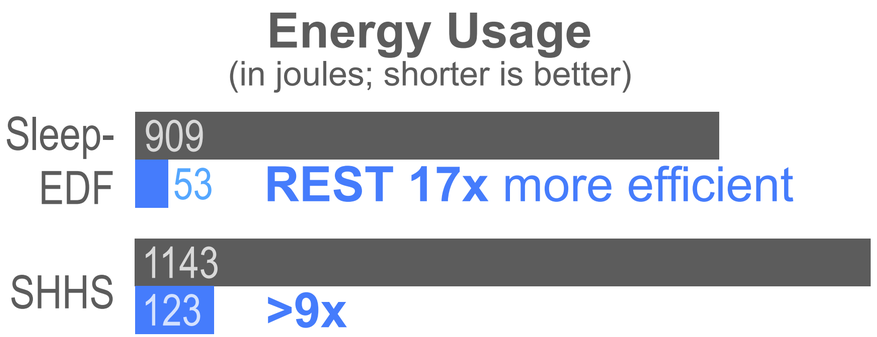


Results: Robustness to Noise



Results: Efficiency

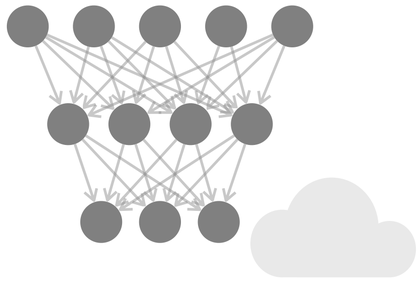




performance evaluated on a pixel 2 smart phone

Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

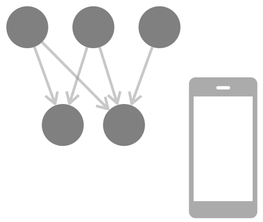
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

HAR: Hardness Aware Reweighing for Imbalanced Datasets

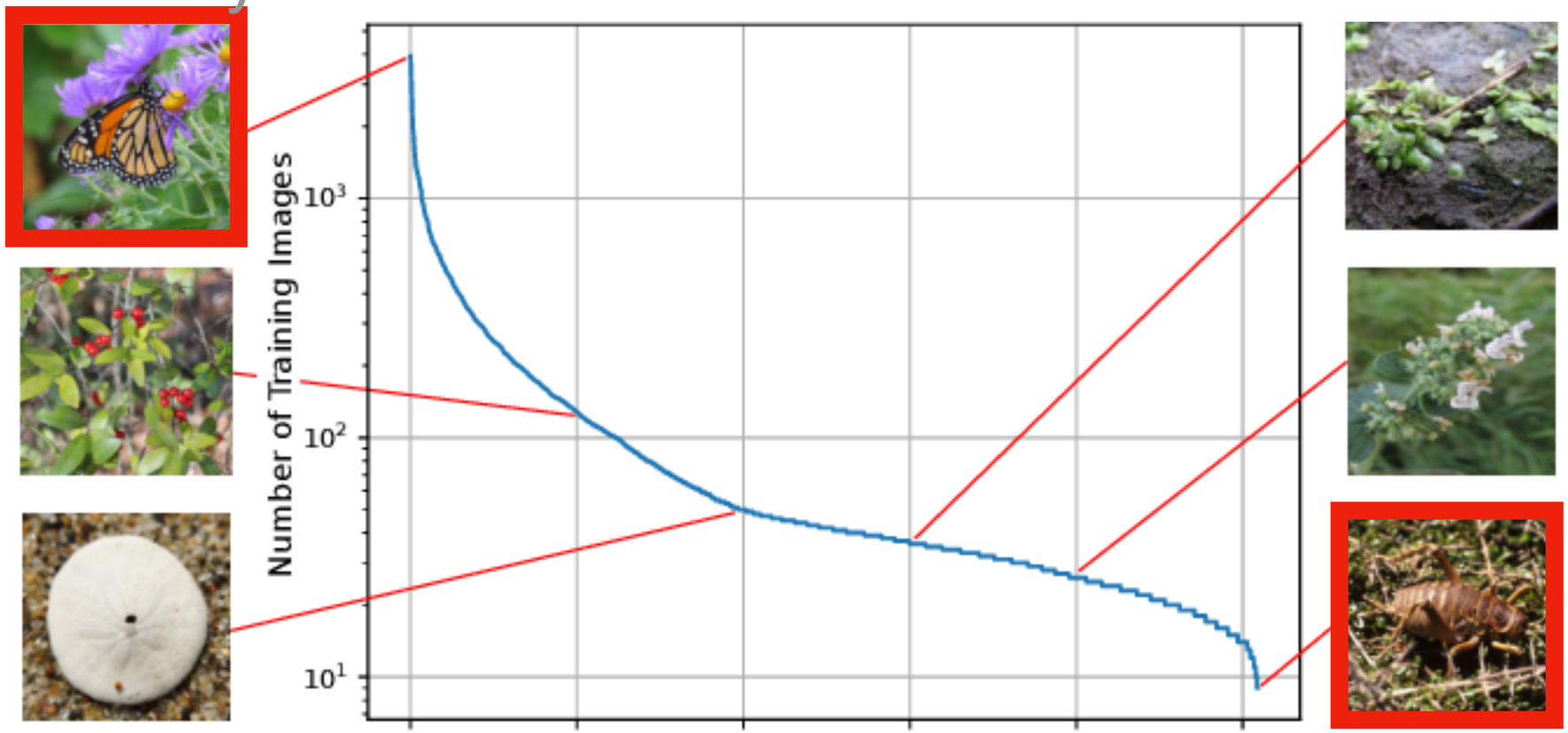
Rahul Duggal, Scott Freitas, Sunny Dhamnani, Duen Horng Chau, Jimeng Sun

IEEE Big Data 2021

Challenging long tail of data

Inaturalist'17 dataset

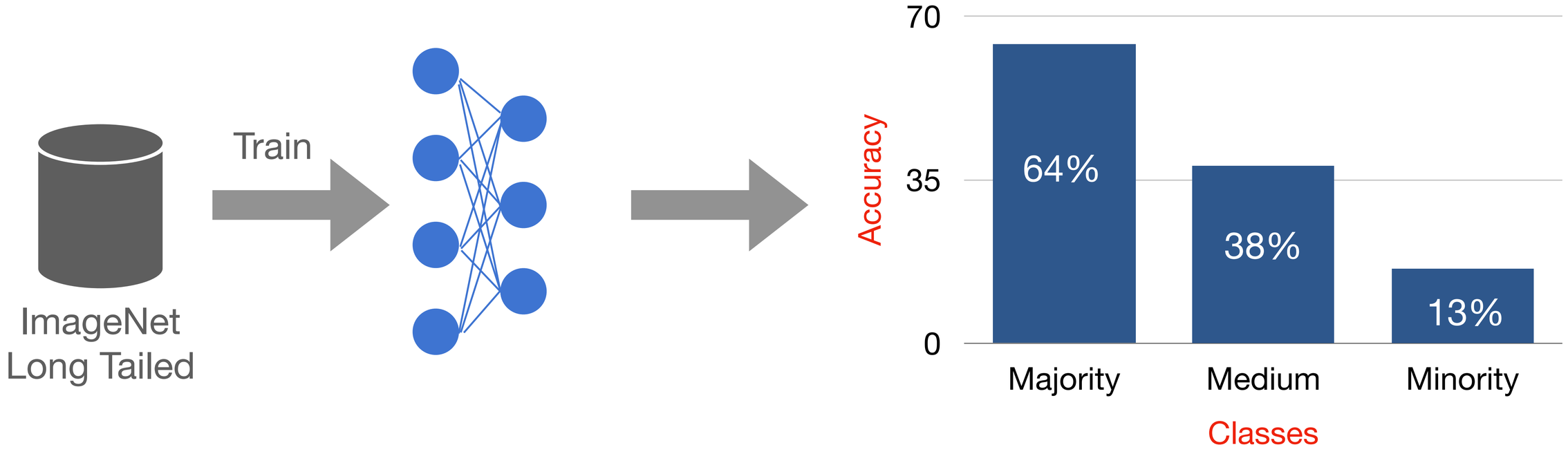
Butterfly



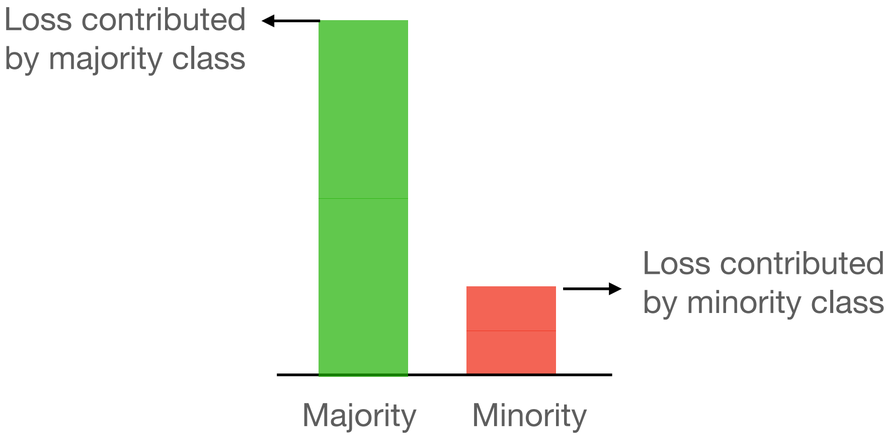
Garden Insect

Horn, Grant Van et al. "The iNaturalist Species Classification and Detection Dataset." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 8769-8778.

Challenging long tail of data

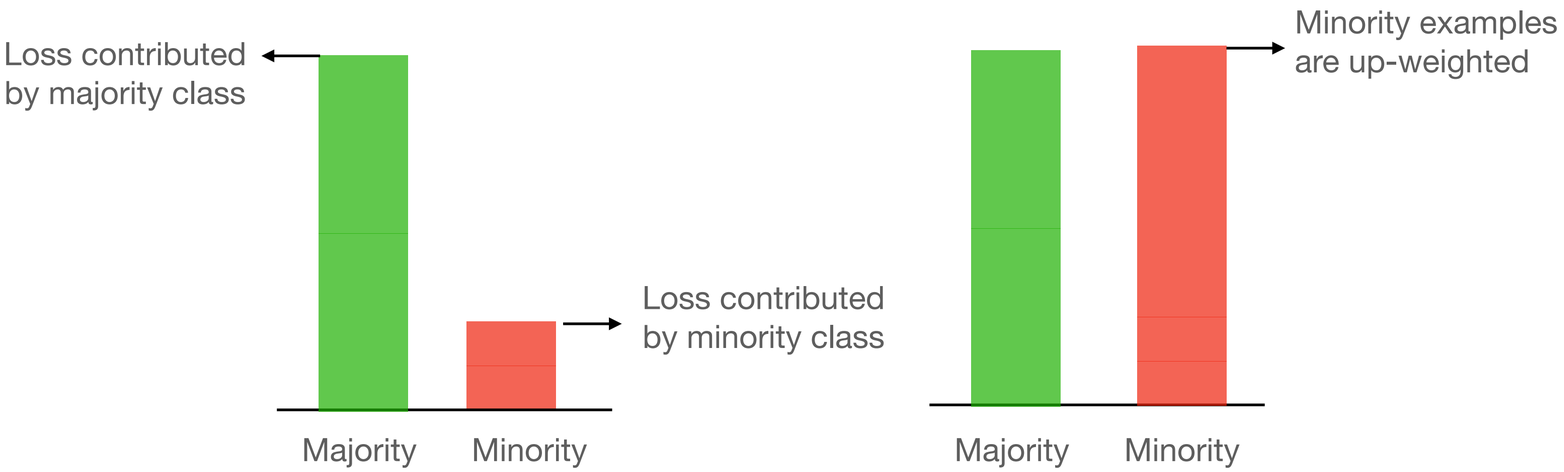


Loss Re-weighting



Original

Loss Re-weighting

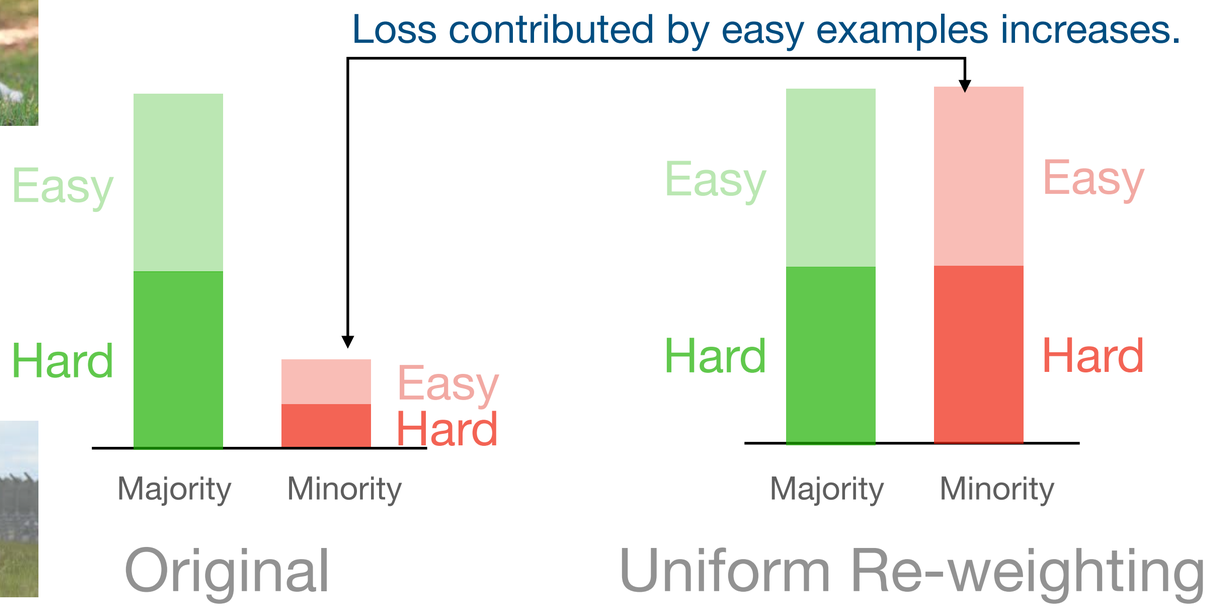


Original

Uniform Re-weighting

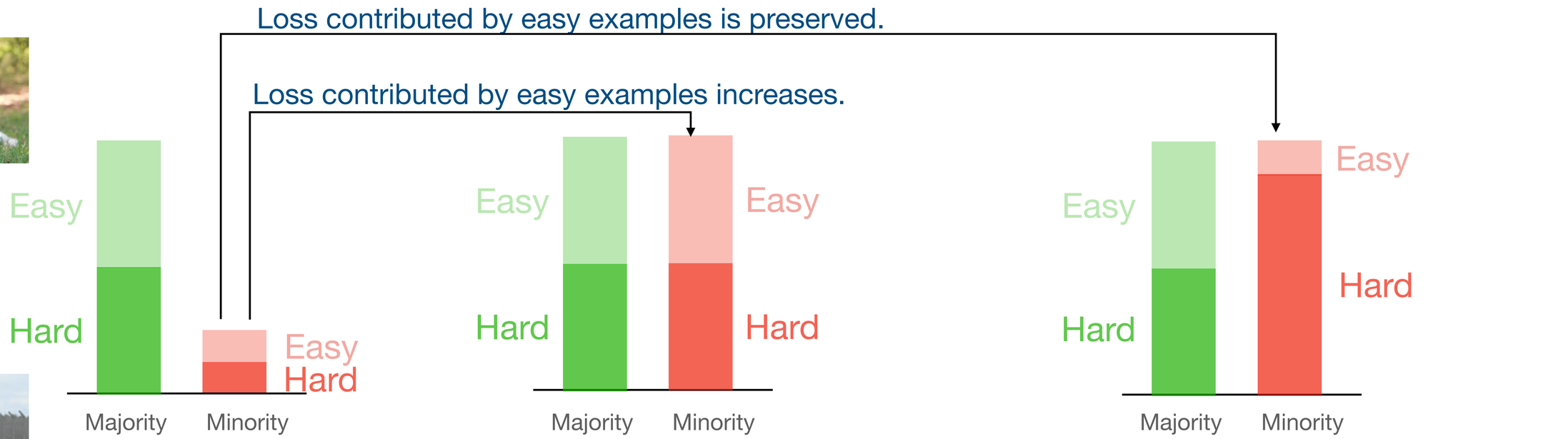
Loss Re-weighting

Loss Re-weighting



Easy and hard examples
up-weighted **equally**

Loss Re-weighting



Original

Uniform Re-weighting

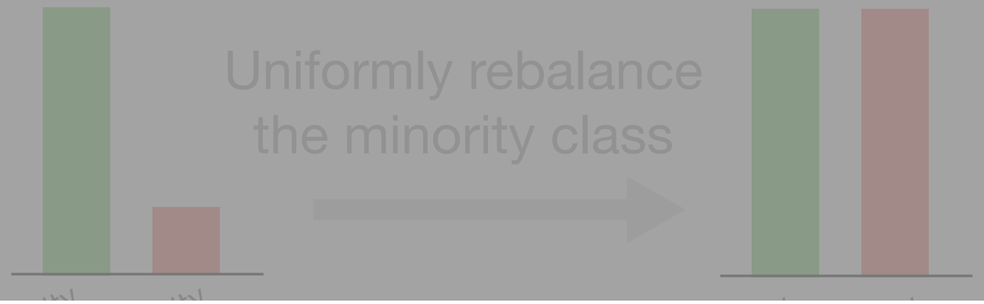
Hardness Aware Reweighting
(ours)

Easy and hard examples up-weighted **equally**

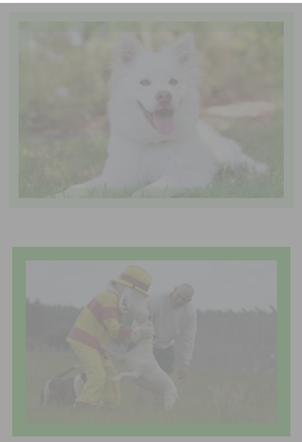
Hard examples up-weighted **more than** easy examples

Mitigating the effect of Imbalance

Class Rebalancing



Hardness Aware Reweighing (HAR) strategy that up-weights only the **hard examples!**



Challenge

No labels for hardness!

Dogs

Easy?

Hard?

Cats

Easy?

Hard?

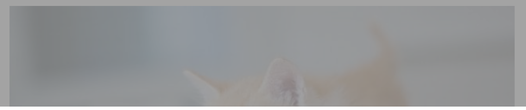
Challenge

No labels for hardness!

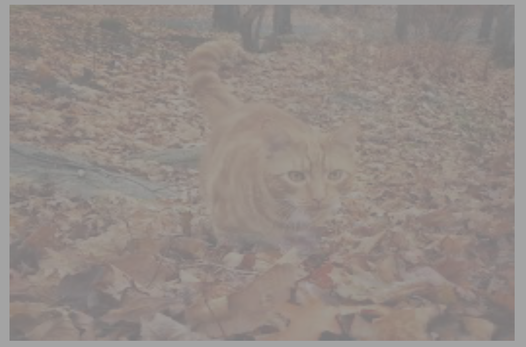
Dogs

Hard?

Cats



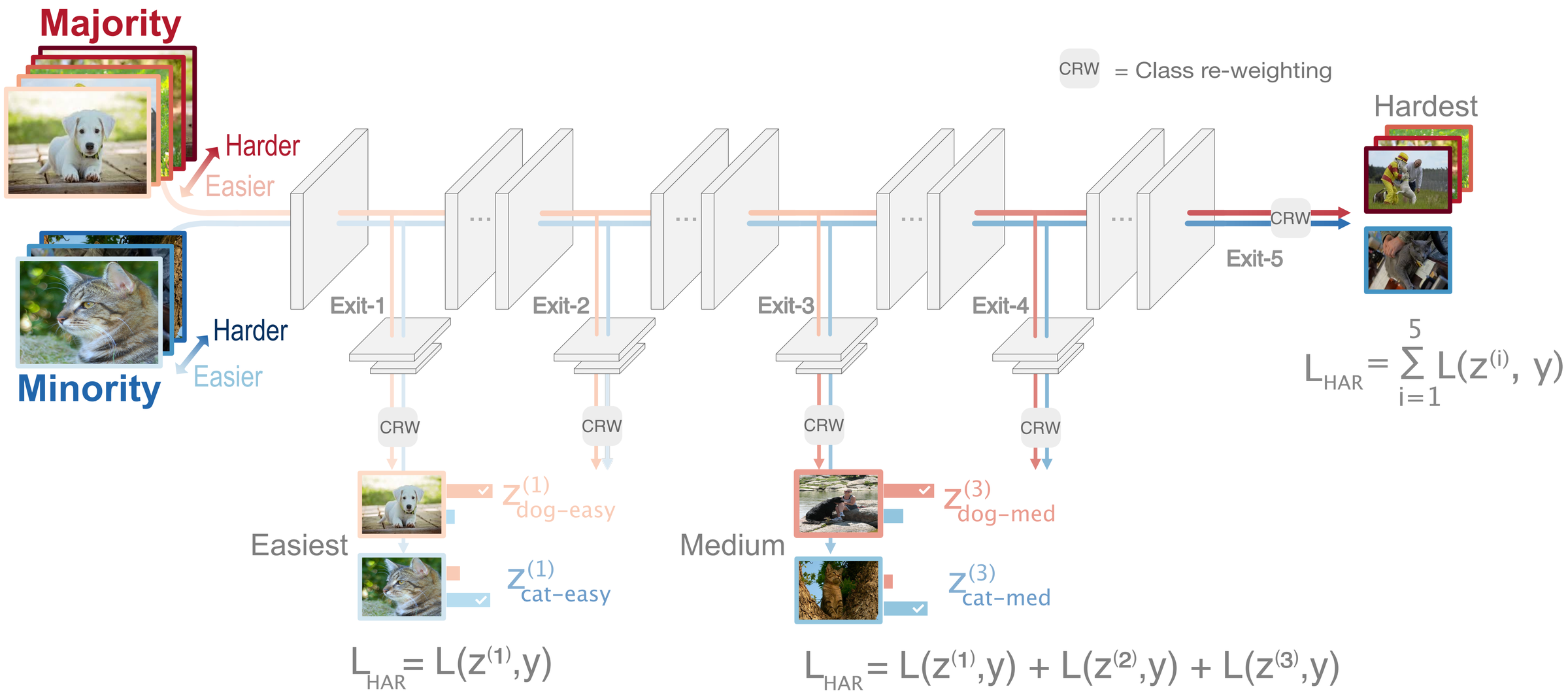
Hard?



How to learn the notion of hardness?

Our method

HAR Framework



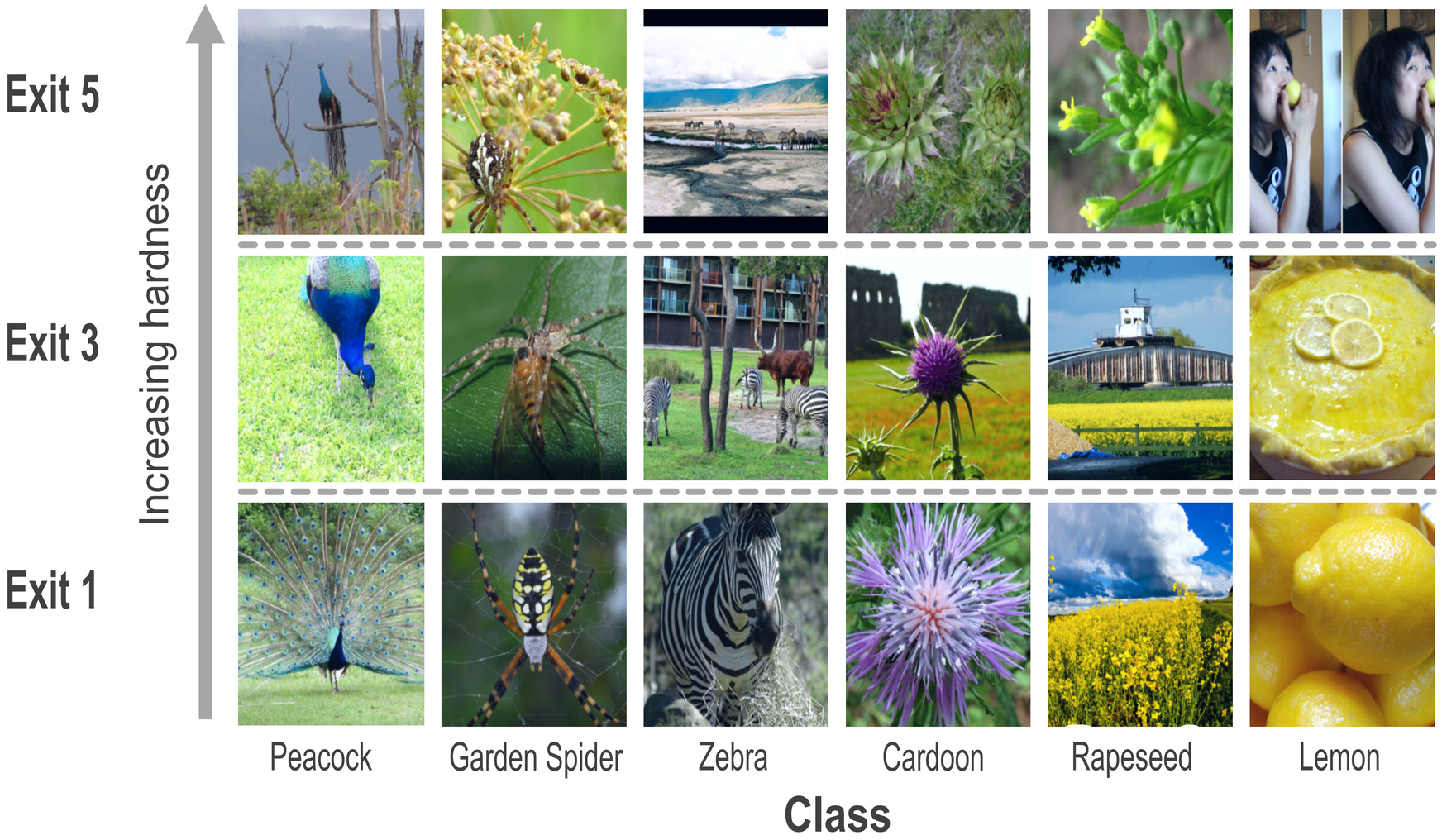
$$L_{HAR} = \sum_{i=1}^{5} L(z^{(i)}, y)$$

$$L_{HAR} = L(z^{(1)}, y)$$

$$L_{HAR} = L(z^{(1)}, y) + L(z^{(2)}, y) + L(z^{(3)}, y)$$

Results - 1

Does HAR learn the notion of hardness?



Results - 2

How does HAR compare to SOTA?

Accuracy Improvement

	CIFAR-10 LT			CIFAR-100 LT		
	100×	50×	10×	100×	50×	10×
CE†	70.4	74.8	83.6	28.3	43.9	55.7
Focal [21]††	70.4	76.7	86.7	28.4	44.3	55.8
Mixup [48]††	73.1	77.8	87.1	39.5	45.0	58.0
Manifold Mixup [49]††	73.0	78.0	87.0	38.3	43.1	56.5
CE+DRW [5]†	76.3	80.0	87.6	41.4	46.0	58.3
HAR(CE)+DRW (Our)	**76.8**	**80.8**	**87.6**	**42.5**	**47.1**	**58.7**
Flops saving	32%	29%	26%	11%	11%	10%
LDAM+DRW [20]†	77.0	81.4	87.6	42.0	46.6	58.7
HAR(LDAM)+DRW (Our)	**78.1**	**82.4**	**88.0**	**43.1**	**47.5**	**58.9**
Flops Saving	15%	21%	20%	0%	2%	3%

Results - 2

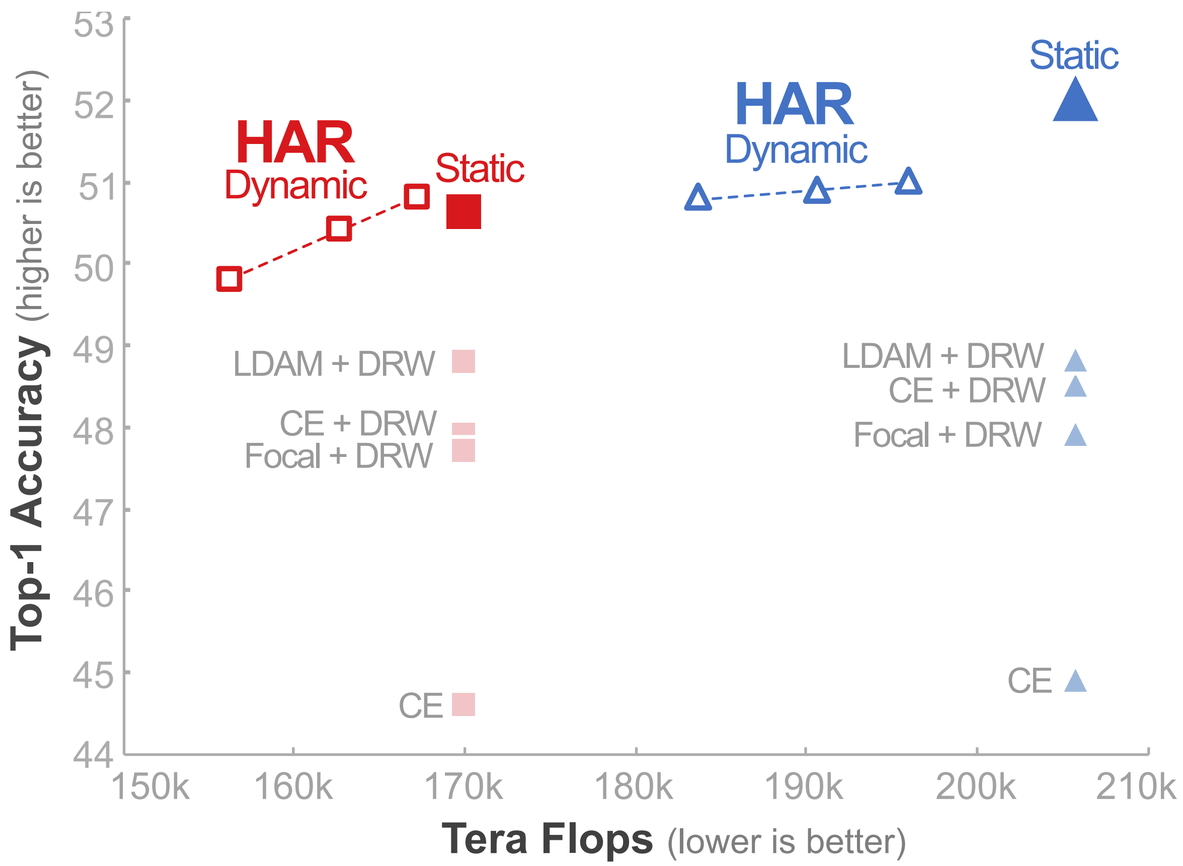
How does HAR compare to SOTA?

Accuracy Improvement

	CIFAR-10 LT			CIFAR-100 LT		
	100×	50×	10×	100×	50×	10×
CE†	70.4	74.8	83.6	28.3	43.9	55.7
Focal [21]††	70.4	76.7	86.7	28.4	44.3	55.8
Mixup [48]††	73.1	77.8	87.1	39.5	45.0	58.0
Manifold Mixup [49]††	73.0	78.0	87.0	38.3	43.1	56.5
CE+DRW [5]†	76.3	80.0	87.6	41.4	46.0	58.3
HAR(CE)+DRW (Our)	**76.8**	**80.8**	**87.6**	**42.5**	**47.1**	**58.7**
Flops saving	32%	29%	26%	11%	11%	10%
LDAM+DRW [20]†	77.0	81.4	87.6	42.0	46.6	58.7
HAR(LDAM)+DRW (Our)	**78.1**	**82.4**	**88.0**	**43.1**	**47.5**	**58.9**
Flops Saving	15%	21%	20%	0%	2%	3%

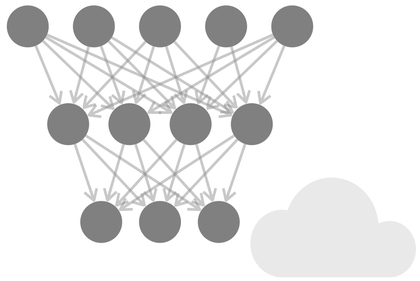
Efficiency Improvement

ResNet50 & DenseNet169



Thesis Overview

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

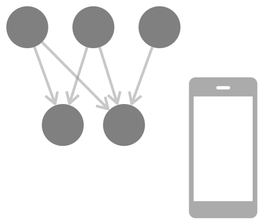
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

IMB-NAS: Neural Architecture Search for Imbalanced Datasets

Rahul Duggal, Sheng-Yun Peng, Hao Zhou, Duen Horng Chau

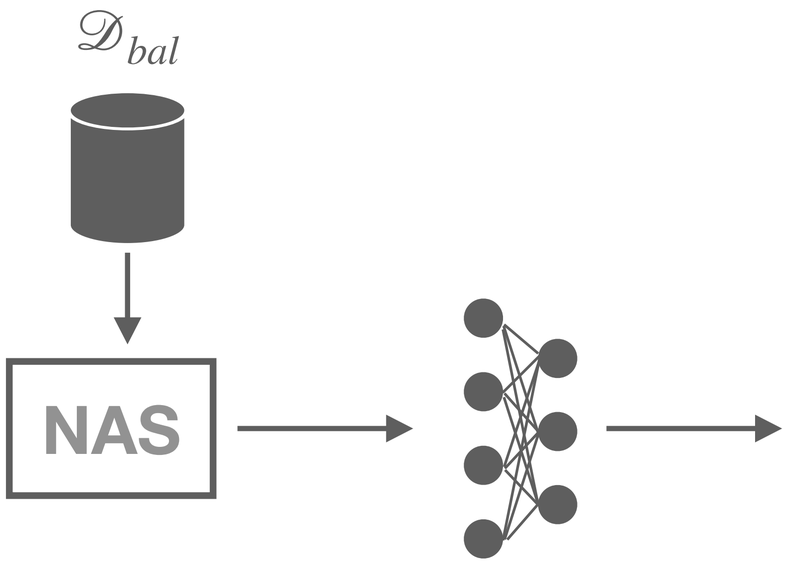
Submitted to AAAI 2023

Motivation

- Neural architecture search (NAS) is the prevalent tool for designing efficient models.

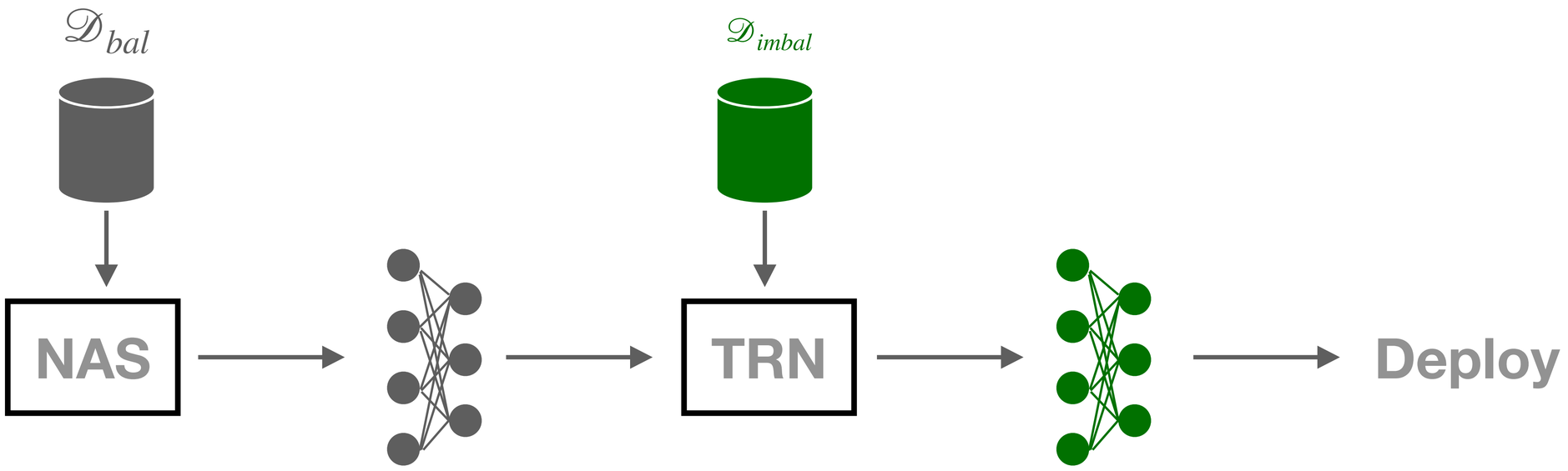
Motivation

- Neural architecture search (NAS) is the prevalent tool for designing efficient models.
- A typical NAS workflow proceeds as follows:



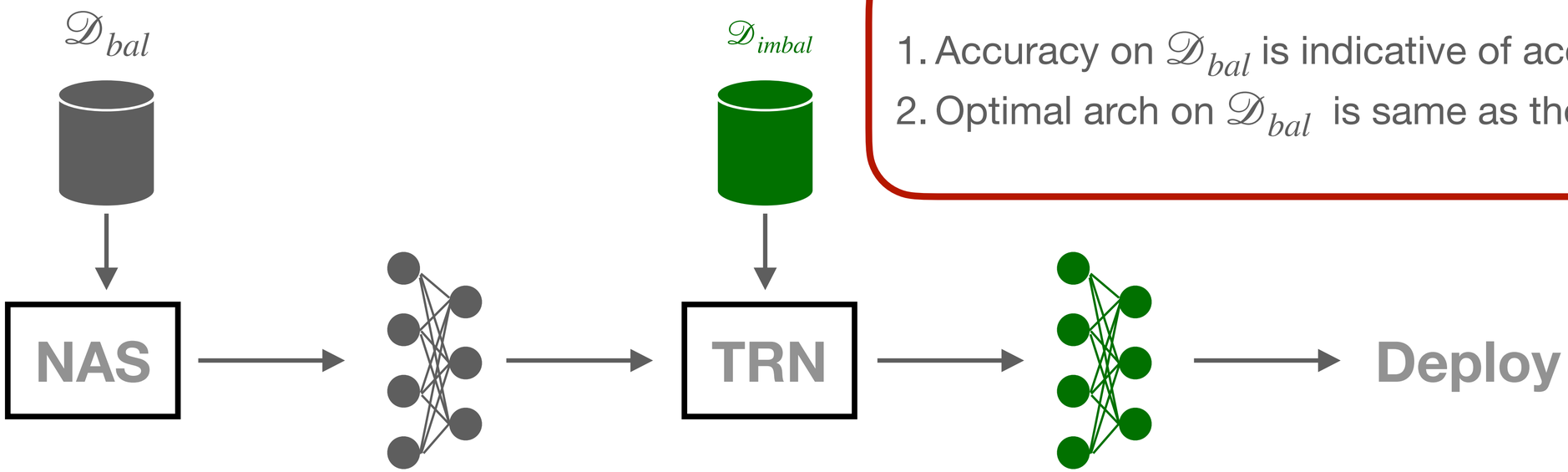
Motivation

- Neural architecture search (NAS) is the prevalent tool for designing efficient models.
- A typical NAS workflow proceeds as follows:



Motivation

- Neural architecture search (NAS) is the prevalent tool for designing efficient models.
- A typical NAS workflow proceeds as follows:



Critical Assumptions

1. Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$.
2. Optimal arch on $\mathscr{D}_{bal}$ is same as the optimal one on $\mathscr{D}_{imbal}$.

Testing critical assumptions

#1 Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$?

$\mathscr{A}$ ~ A_1 A_2

DARTS search space

Testing critical assumptions

#1 Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$?

$\mathscr{A}$ ~ A_1, A_2

DARTS search space

Dataset	Model	Flops	Accuracy (%) bal(1×)	Accuracy (%) imbal(100×)
Cifar10	A1	410	94.6	77.3
	A2	407	94.7	74.1

A_1 outperforms A_2 by 3% on $\mathscr{D}_{imbal}$

A_2 outperforms A_1 by 0.1% on $\mathscr{D}_{bal}$

Testing critical assumptions

#1 Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$?

$\mathscr{A}$ ~ A_1, A_2

DARTS search space

Dataset	Model	Flops	Accuracy (%) bal(1×)	Accuracy (%) imbal(100×)
Cifar10	A1	410	94.6	77.3
	A2	407	94.7	74.1

A_1 outperforms A_2 by 3% on $\mathscr{D}_{imbal}$

A_2 outperforms A_1 by 0.1% on $\mathscr{D}_{bal}$

#2 Optimal arch on $\mathscr{D}_{bal}$ is same as the optimal one on $\mathscr{D}_{imbal}$?

$\mathscr{A}$ ~ A_1 ⋮ A_{134}

NATS search space

Testing critical assumptions

#1 Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$?

DARTS search space

$\sim$ A_1, A_2

Dataset	Model	Flops	Accuracy (%) bal(1×)	Accuracy (%) imbal(100×)
Cifar10	A1	410	94.6	77.3
	A2	407	94.7	74.1

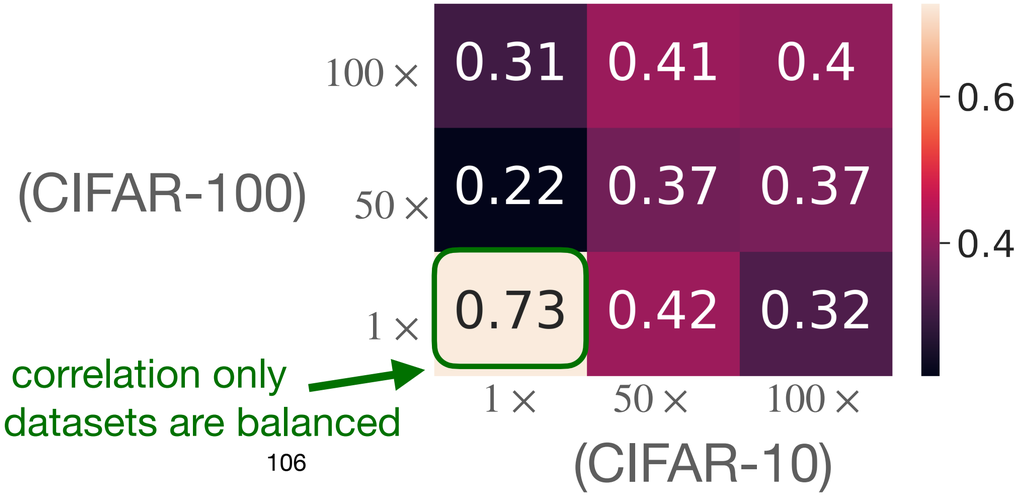
A_1 outperforms A_2 by 3% on $\mathscr{D}_{imbal}$

A_2 outperforms A_1 by 0.1% on $\mathscr{D}_{bal}$

#2 Optimal arch on $\mathscr{D}_{bal}$ is same as the optimal one on $\mathscr{D}_{imbal}$?

NATS search space

$\sim$ A_1 ⋮ A_{134}



High correlation only when both datasets are balanced

Testing critical assumptions

#1 Accuracy on $\mathscr{D}_{bal}$ is indicative of accuracy on $\mathscr{D}_{imbal}$?

$\mathscr{A}$

DARTS search space

~ A_1 A_2

Dataset	Model	Flops	Accuracy (%) bal(1×)	Accuracy (%) imbal(100×)
Cifar10	A1	410	94.6	77.3
	A2	407	94.7	74.1

A_1 outperforms A_2 by 3% on $\mathscr{D}_{imbal}$

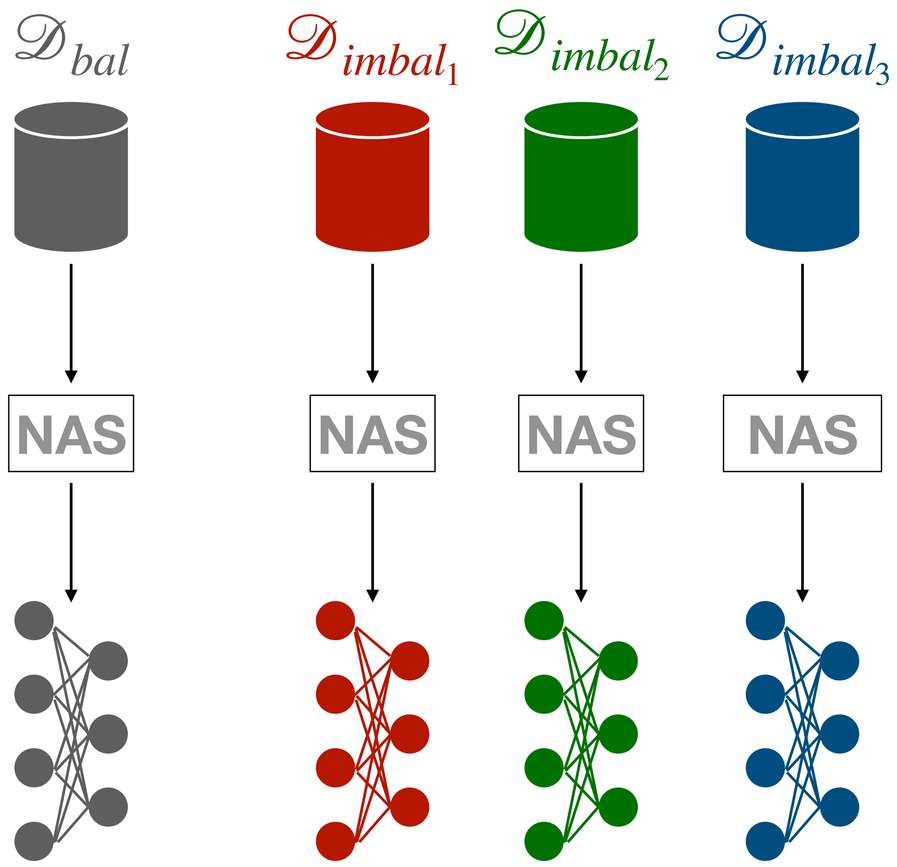
A_2 outperforms A_1 by 1% on $\mathscr{D}_{bal}$

#2 Optimal arch on $\mathscr{D}_{bal}$ is same as the optimal one on $\mathscr{D}_{imbal}$?

Architecture ranking on imbalanced datasets differs significantly from the ranking on balanced datasets.

NAS for Imbalanced Datasets

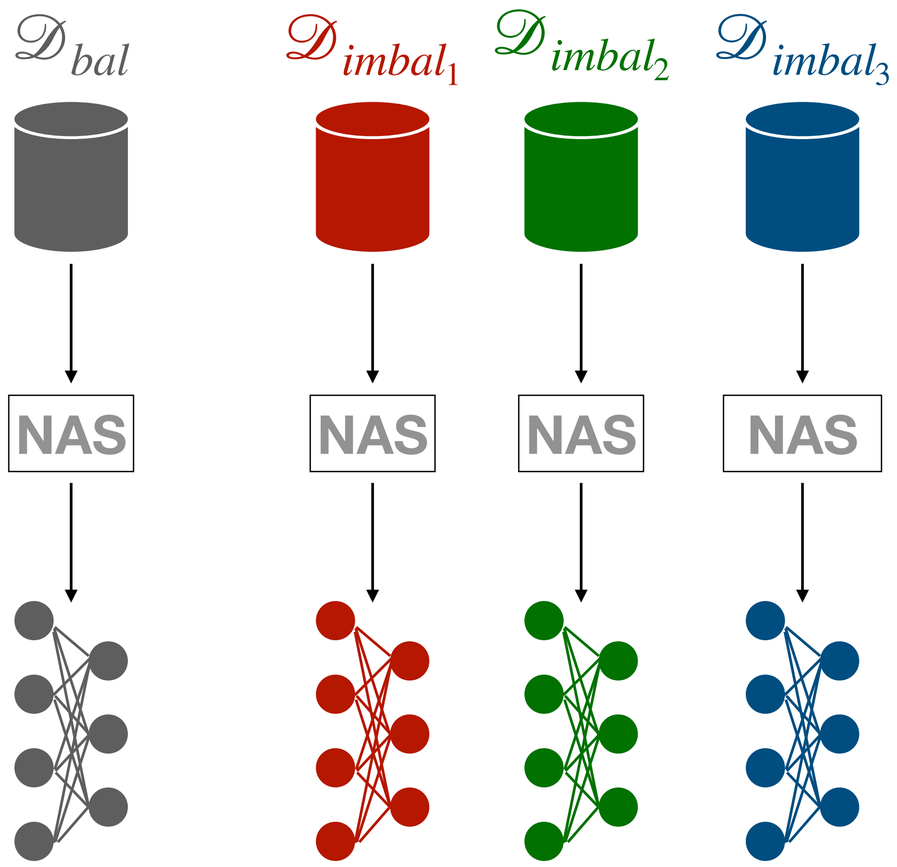
Existing NAS workflow



Very compute intensive!

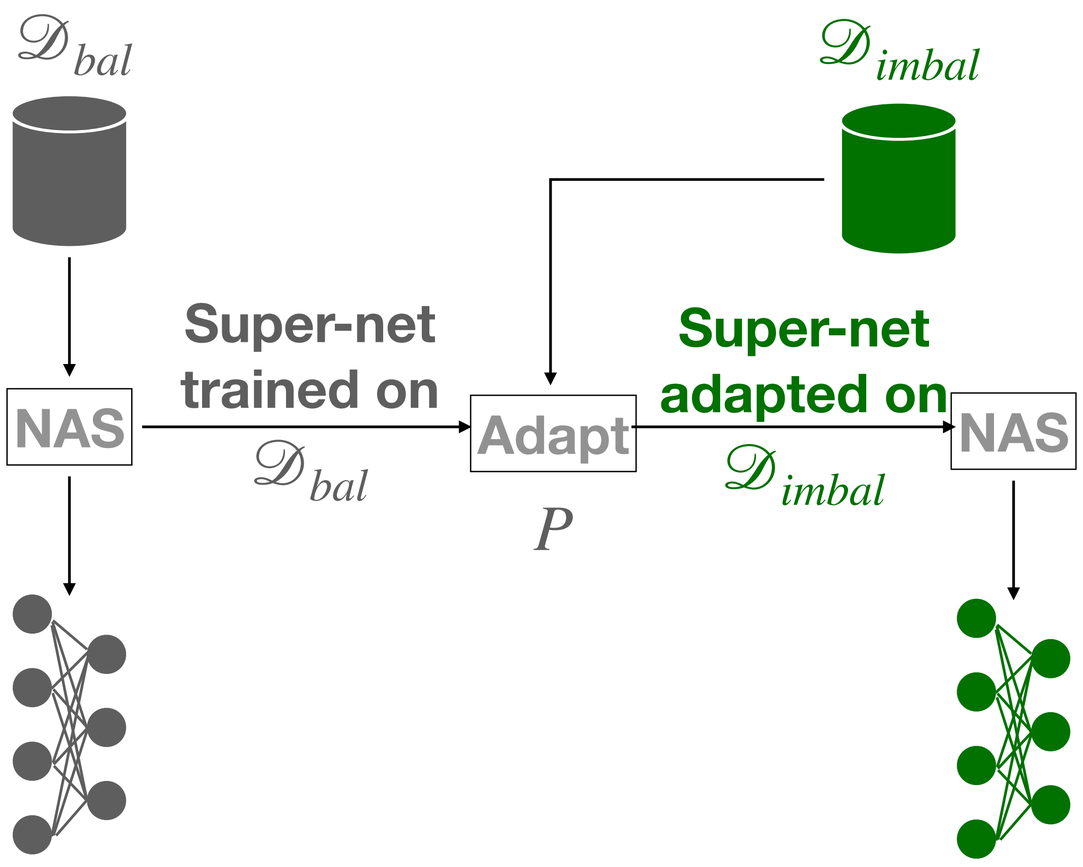
NAS for Imbalanced Datasets

Existing NAS workflow



Very compute intensive!

Proposed workflow

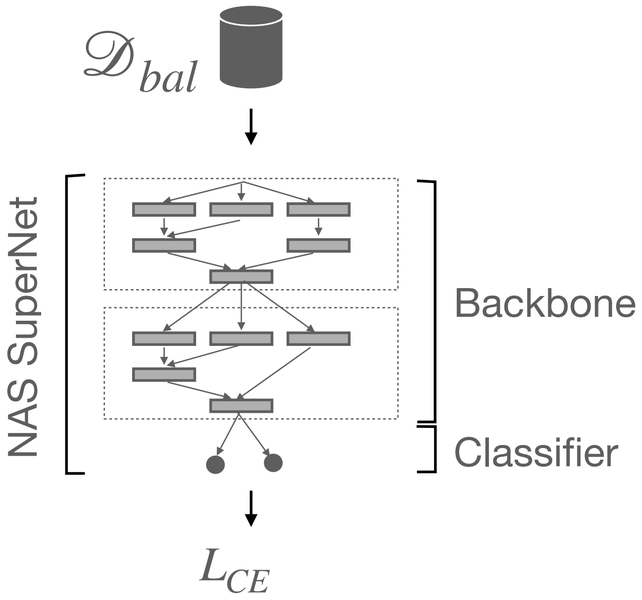


Highly efficient

NAS for Imbalanced Datasets

Super-network Adaptation Strategies

No Adaption

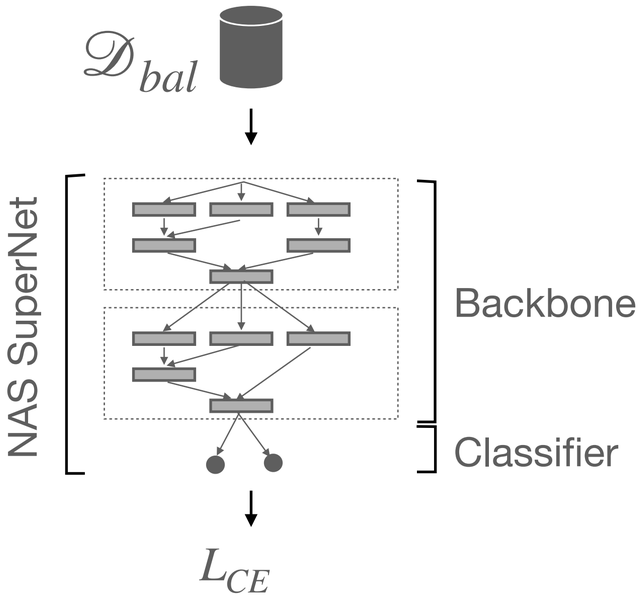


$$w^*_{a,\mathscr{D}_s}, w^*_{c,\mathscr{D}_s} = min_{w_a,w_c} \mathbb{E}_{a\sim\mathscr{A}}(L_{CE}(\phi(w_c), \phi(a, w_a); \mathscr{D}_s)$$

NAS for Imbalanced Datasets

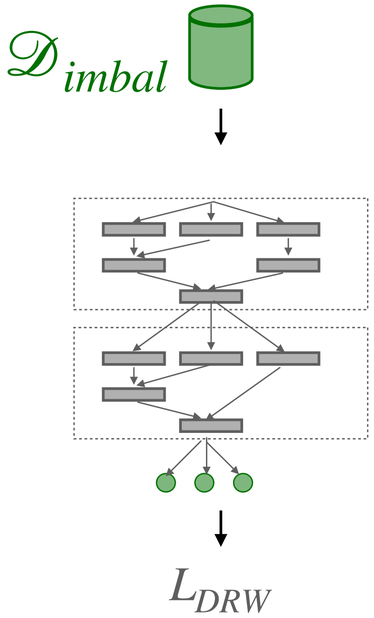
Super-network Adaptation Strategies

No Adaption



$$w^*_{a,\mathscr{D}_s}, w^*_{c,\mathscr{D}_s} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathscr{A}}(L_{CE}(\phi(w_c), \phi(a, w_a); \mathscr{D}_s)$$

Adapt Classifier

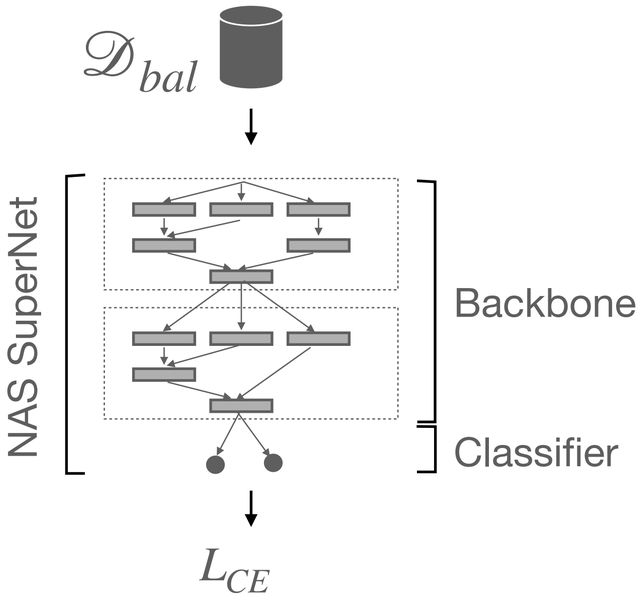


$$w^*_{c,\mathscr{D}_t} = min\mathbb{E}_{a\sim\mathscr{A}}(L_{DRW}(\phi(w_c), \phi(a, w^*_{a,\mathscr{D}_s}); \mathscr{D}_t)$$

NAS for Imbalanced Datasets

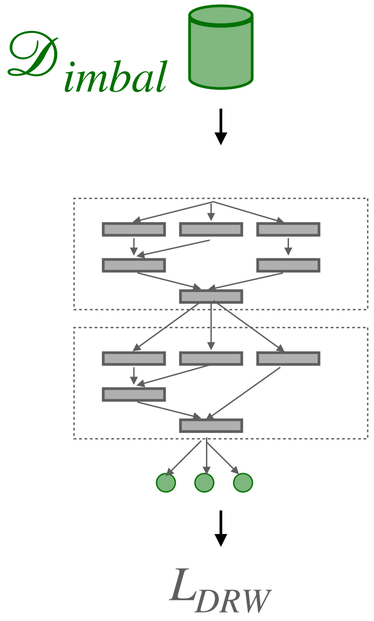
Super-network Adaptation Strategies

No Adaption



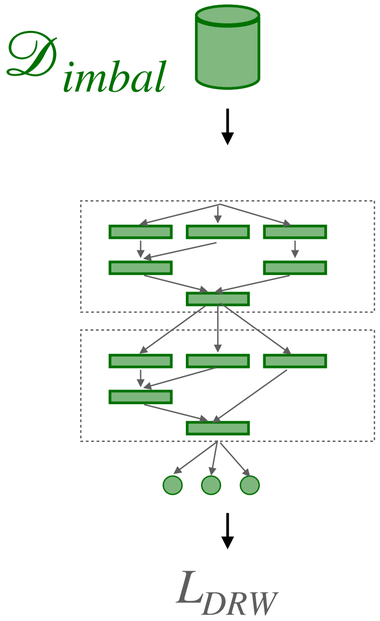
$$w^*_{a,\mathscr{D}_s}, w^*_{c,\mathscr{D}_s} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathscr{A}}(L_{CE}(\phi(w_c), \phi(a, w_a); \mathscr{D}_s)$$

Adapt Classifier



$$w^*_{c,\mathscr{D}_t} = min\mathbb{E}_{a\sim\mathscr{A}}(L_{DRW}(\phi(w_c), \phi(a, w^*_{a,\mathscr{D}_s}); \mathscr{D}_t)$$

Adapt Backbone+Classifier

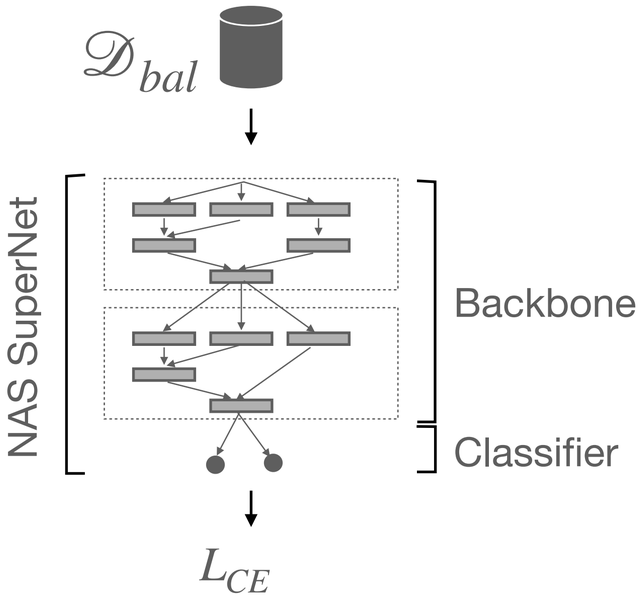


$$w^{**}_{a,\mathscr{D}_t}, w^*_{c,\mathscr{D}_t} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathscr{A}}(L_{DRW}(\phi(w_c), \phi(a, w^*_{a,\mathscr{D}_s}); \mathscr{D}_t)$$

NAS for Imbalanced Datasets

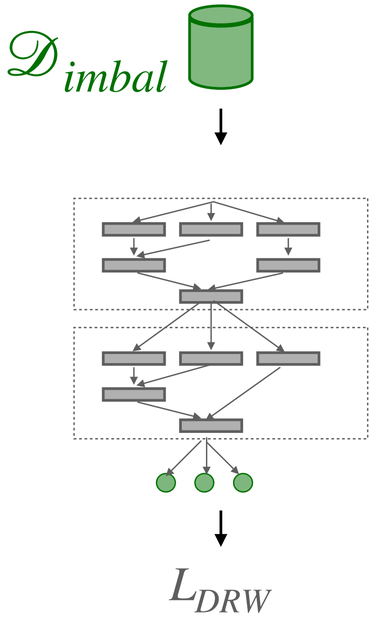
Super-network Adaptation Strategies

No Adaption



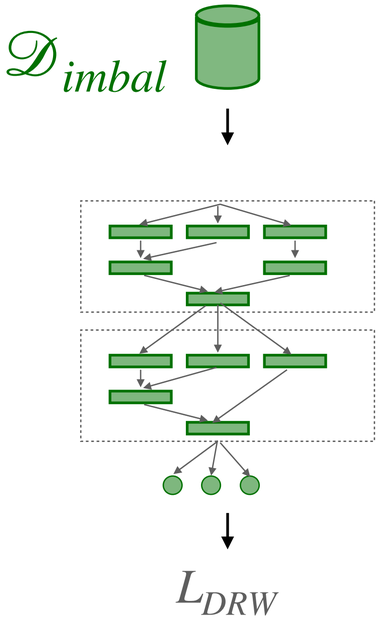
$$w^*_{a,\mathcal{D}_s}, w^*_{c,\mathcal{D}_s} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathcal{A}}(L_{CE}(\phi(w_c),\phi(a,w_a);\mathcal{D}_s)$$

Adapt Classifier



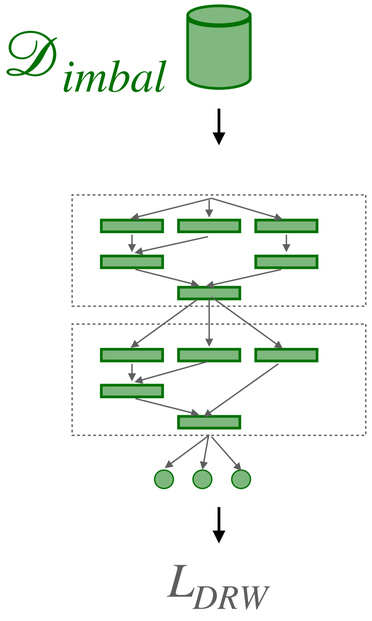
$$w^*_{c,\mathcal{D}_t} = min\mathbb{E}_{a\sim\mathcal{A}}(L_{DRW}(\phi(w_c),\phi(a,w^*_{a,\mathcal{D}_s});\mathcal{D}_t)$$

Adapt Backbone+Classifier



$$w^{**}_{a,\mathcal{D}_t}, w^*_{c,\mathcal{D}_t} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathcal{A}}(L_{DRW}(\phi(w_c),\phi(a,w^*_{a,\mathcal{D}_s});\mathcal{D}_t)$$

Training backbone & classifier From Scratch



$$w^*_{a,\mathcal{D}_t}, w^*_{c,\mathcal{D}_t} = min_{w_a,w_c}\mathbb{E}_{a\sim\mathcal{A}}(L_{CE}(\phi(w_c),\phi(a,w_a);\mathcal{D}_t)$$

Results-1 (Accuracy)

	CIFAR-10-balance → CIFAR-100-imbalance				CIFAR-100-balance → CIFAR-10-imbalance			
	Imbalance Ratio				Imbalance Ratio			
	50×	100×	200×	400×	100×	200×	400×	800×
No Adaption (baseline)	45.80	40.83	36.30	32.80	75.96	68.96	63.26	56.90
From Scratch (paragon)	45.93	41.53	37.03	33.40	76.03	70.23	63.96	57.70

Observations

- Paragon always outperforms baseline.

Results-1 (Accuracy)

	CIFAR-10-balance → CIFAR-100-imbalance				CIFAR-100-balance → CIFAR-10-imbalance			
	Imbalance Ratio				Imbalance Ratio			
	50×	100×	200×	400×	100×	200×	400×	800×
No Adaption (baseline)	45.80	40.83	36.30	32.80	75.96	68.96	63.26	56.90
Classifier (ours)	**45.06**	**41.93**	**36.76**	**33.70**	**75.93**	**69.70**	**63.80**	**58.23**
Backbone+Classifier (ours)	44.86	41.86	36.70	33.46	75.86	69.26	63.70	58.03
From Scratch (paragon)	45.93	41.53	37.03	33.40	76.03	70.23	63.96	57.70

Observations

- Paragon always outperforms baseline.
- Both **our adaptation procedures** (shaded yellow) outperform the baseline.
- **Classifier-only** adaptation (bolded) outperforms **backbone+classifier** adaptation.

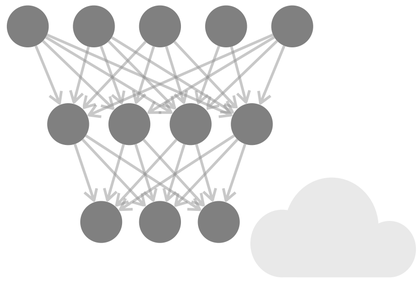
Results-2 (Compute Cost)

Conclusion

Classifier-only adaptation is more efficient and effective

Conclusions and Impact

Robust Efficient Edge AI



Large models deployed on cloud servers

Part I: New framework for **efficient edge AI**

★ **CMP-NAS**
Bottom-up via NAS
[CVPR 21]

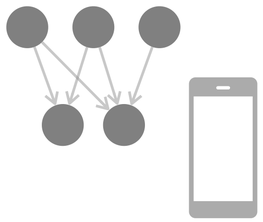
★ **CUP**
Top-down via pruning
[Big Data 21]

Part II: New methods for **robust edge AI**

★ **REST**
Test-time robustness via pruning [WWW 20]

★ **HAR**
Training-time robustness via early-exit
[Big Data 21]

★ **IMB-NAS**
Training-time robustness via NAS [Submitted to AAAI 2023]



Small models deployed on edge devices

Thesis Statement

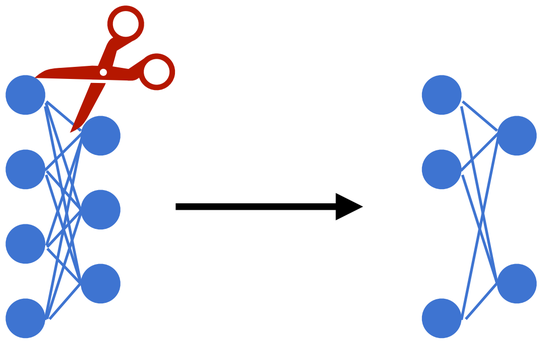
My thesis develops new **algorithmic frameworks** that empower AI on edge devices by making deep neural networks **efficient** and **robust** to noise.

Dissertation Contributions

New algorithms across areas

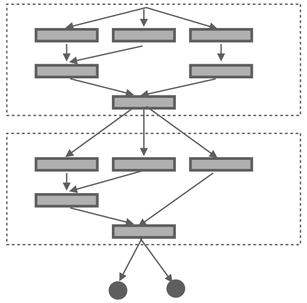
1. **Neural Network Pruning**

   CUP, REST



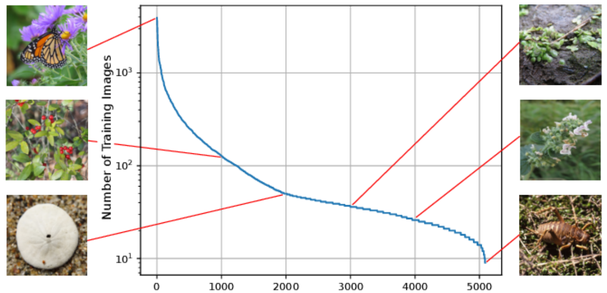
2. **Neural Architecture Search**

   CMP-NAS, IMB-NAS



3. **Long Tail Imbalance**

   HAR



Dissertation Impact: Real-world Applications

CMP-NAS: First system to extend image-based product search onto edge devices. Highlighted by AWS CEO. Entering production.

Adam Selipsky
@aselipsky

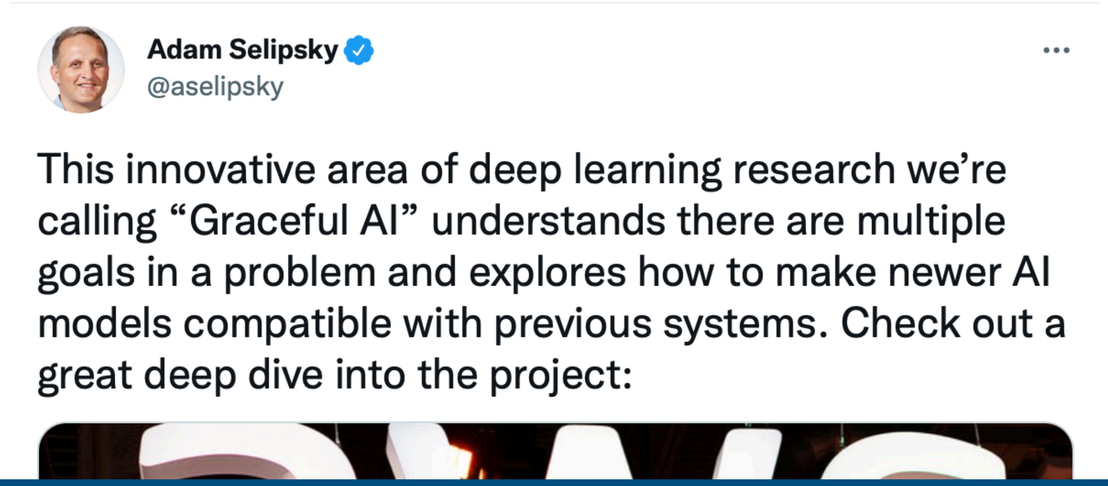
This innovative area of deep learning research we're calling "Graceful AI" understands there are multiple goals in a problem and explores how to make newer AI models compatible with previous systems. Check out a great deep dive into the project:

REST: First to demonstrate viability of home-based sleep diagnosis. Featured in news media.

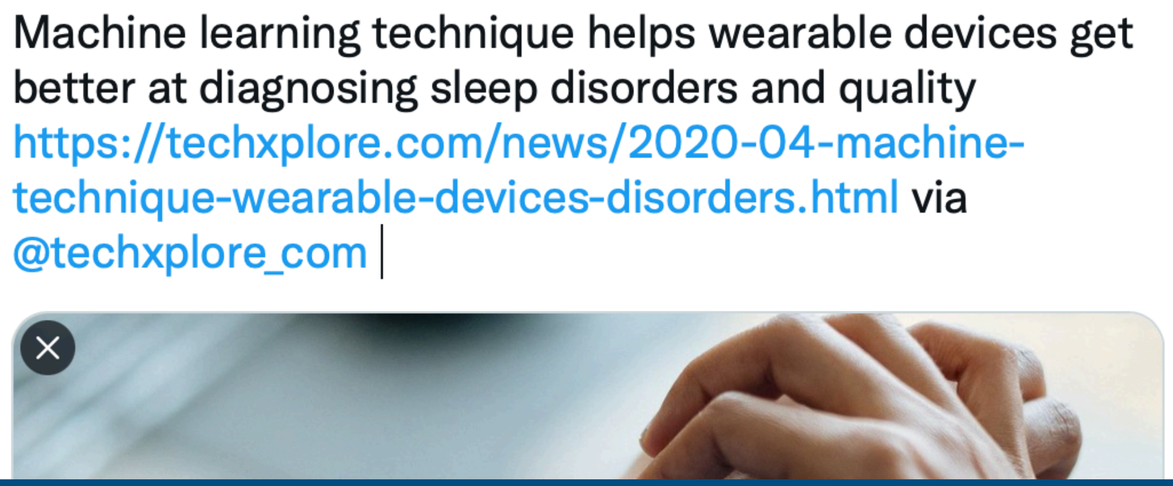
Machine learning technique helps wearable devices get better at diagnosing sleep disorders and quality https://techxplore.com/news/2020-04-machine-technique-wearable-devices-disorders.html via @techxplore_com

Dissertation Impact: Real-world Applications

CMP-NAS: First system to extend image-based product search onto edge devices. Highlighted by AWS CEO. Entering production.



REST: First to demonstrate viability of home-based sleep diagnosis. Featured in news media.



Amazon post internship Fellowship

Post PhD…

Thanks!

ROBUST EFFICIENT EDGE AI:
New Principles and Frameworks for Empowering AI on Edge

Polo Chau, Jimeng Sun, Richard Vuduc, Sunny Dhamnani, Karan Samel, Monica Isgut, Irfan Al Husseini, Tianfan Fu, Siddharth Biswal, Yanbo Xu, Devyani Choudhary, Scott Freitas, Haekyu Park, Nilaksh Das, Fred Hohman, Zijie Wang, Austin Wright, Ben Hoover, Seongmin Lee, Pramod Chunduri, Aditya Venkatraman, Jyoti Narang, Hao Zhou, Shuo Yang, Yuanjun Xiong, Yue Zhao, Zhuowen Tu, Wei Xia, Stefano Soatto, Gaurav Duggal, Manisha Duggal, Vijay Duggal, Madhvi Dua.

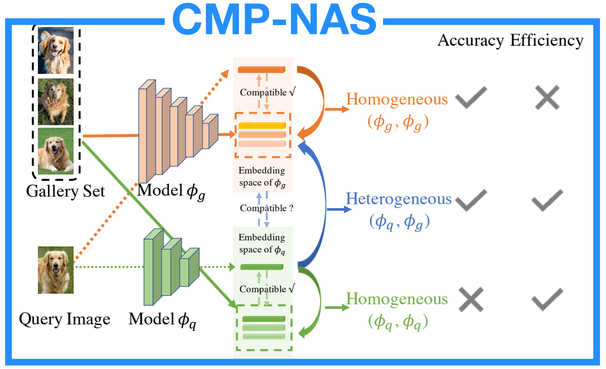
Polo club of data science

Sunlab

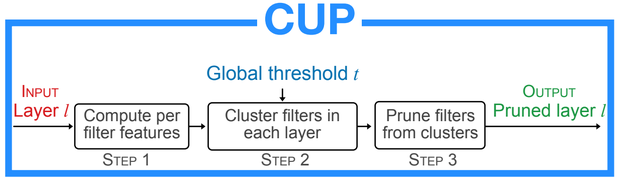
Amazon AI (Computer Vision Group)

Georgia Tech College of Computing
Computational Science and Engineering

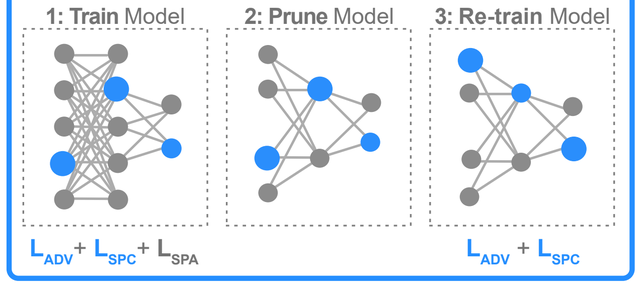
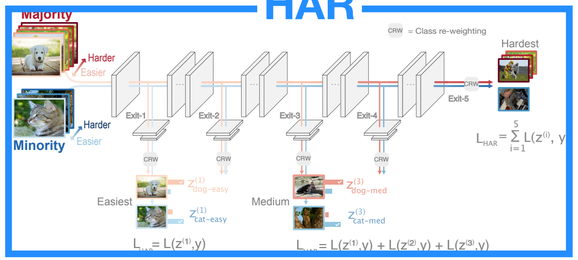
CMP-NAS



CUP



REST-PROCESS



HAR